

# Companion to the Nursing Ph.D. & DNP Statistics Curricula

Wm. Ellery Samuels, Ph.D.

September 25, 2025



# Table of contents

<b>Preface</b>	<b>1</b>
<b>I Applied Statistics Curriculum</b>	<b>3</b>
<b>1 NURS 60N: Foundations of Biostatistics for Nursing Research and Evidence-Based Practice</b>	<b>9</b>
1.1 Major Principles . . . . .	9
1.2 Frequencies & Counts . . . . .	9
1.3 Measuring & Testing Differences . . . . .	10
1.4 Power and Effect Size . . . . .	10
1.5 Association & Causation . . . . .	10
1.6 The ANOVA Family of Tests . . . . .	10
<b>2 NURS 915 &amp; 916: Applied Statistics 1 &amp; 2</b>	<b>13</b>
2.1 Review of Some Major Principles & Practices in Biostatistics . . . . .	13
2.2 Presenting Results . . . . .	13
2.2.1 Writing Results . . . . .	13
2.3 Introduction to Linear Models . . . . .	14
2.4 Testing Models Theoretically . . . . .	14
2.5 Analyses of Longitudinal Data . . . . .	15
2.6 Logistic Regresssion . . . . .	15
2.7 Structural Equation Modeling . . . . .	15
<b>3 Writing Results Sections</b>	<b>17</b>
3.1 Overview . . . . .	17
3.2 Setting Global Options in SPSS and Using a Word Template . . . . .	17
3.2.1 Setting Global Options in SPSS . . . . .	17
3.2.2 Using a Word Template . . . . .	18

3.3	Data Prep and Cleaning . . . . .	18
3.3.1	Adding Variable Value Labels . . . . .	19
3.3.2	Changing Measurement Scale . . . . .	19
3.3.3	Creating Standardized Variables . . . . .	19
3.3.4	Creating Dummy Variables . . . . .	21
3.4	Exploring the Data . . . . .	23
3.4.1	Descriptives . . . . .	23
3.4.2	Guided Data Exploration . . . . .	25
3.4.3	Copying a Figure to Word . . . . .	29
3.4.4	Correlations . . . . .	29
3.5	Creating Figures in SPSS . . . . .	31
3.6	Analyzing Predictors of Self-Reported GEC Slope . . . . .	33
3.6.1	First Model . . . . .	33
3.6.2	Second (Final) Model . . . . .	38
3.7	Writing the Results . . . . .	40
3.7.1	Overall Strategy . . . . .	40
3.7.2	Writing Style . . . . .	42
3.8	Writing about These Results . . . . .	44
3.8.1	Descriptives . . . . .	45
3.8.2	Inferentials . . . . .	49
3.9	Further Resources . . . . .	50
<b>4</b>	<b>Introduction to Measuring Relationships and Building Models</b>	<b>53</b>
4.1	Measuring Relationships . . . . .	53
4.2	Signal-to-Noise Ratios . . . . .	54
4.3	Building Models . . . . .	54
<b>5</b>	<b>Variance, Covariance, Correlations, and Partial Correlations</b>	<b>59</b>
5.1	The Roles of Covariance & Variance . . . . .	59
5.2	Correlations . . . . .	61
5.3	Partial & Semipartial Correlations . . . . .	61
5.4	Investigating Why DBT Works or Doesn't Work . . . . .	62
5.4.1	Your Task . . . . .	63
5.4.2	Description of the Data . . . . .	63
5.5	Using SPSS . . . . .	68
5.5.1	Accessing SPSS & Data Importation . . . . .	68
5.5.2	Conducting Correlations and Partial Correlations in SPSS . . . . .	69

<b>6 Effect Size: Explanation and Guidelines</b>	<b>71</b>
6.1 Common Effect Sizer Statistics . . . . .	72
6.1.1 Mean Differences . . . . .	72
6.1.2 Proportions of Variance Explained . . . . .	74
6.1.3 Odds & Risk Ratios . . . . .	77
6.2 “Small,” “Medium,” & “Large” Effects . . . . .	78
6.2.1 Effect Size Criteria as Percent of Total Variance . . . . .	78
6.2.2 Effect Size Criteria as Noticeability of Effects . . . . .	79
6.2.3 Effect Size Criteria for Odds Ratios . . . . .	80
6.2.4 A Few Words of Caution About Effect Size Criteria . . . . .	81
6.2.5 Table of Effect Size Statistics . . . . .	82
6.3 Converting Between Effect Size Measures . . . . .	83
6.3.1 A Few Notes on Conversions . . . . .	84
6.3.2 Cohen’s $f$ (and $f^2$ ) to Cohen’s $d$ . . . . .	84
6.3.3 Cohen’s $d$ and Student’s $t$ . . . . .	85
6.3.4 $\eta^2$ and $F$ -scores . . . . .	85
6.4 Additional Resources . . . . .	85
<b>7 Missing Data</b>	<b>87</b>
7.1 Missing Data: More Than Just a Smaller Sample . . . . .	87
7.1.1 Sources of Attrition . . . . .	87
7.1.2 Addressing Attrition . . . . .	88
<b>8 Linear Regression Modeling with SPSS, Part 1: Introduction</b>	<b>91</b>
8.1 Overview . . . . .	91
8.2 Core Concepts . . . . .	91
8.2.1 Linear Relationships . . . . .	91
8.2.2 Consider Removing the Intercept . . . . .	92
8.3 Introduction to Linear Regression Models . . . . .	92
8.3.1 Correlation vs. Simple Linear Regression . . . . .	92
8.3.2 Conducting a Multivariate Linear Regression Using Forward Term Selection . . . . .	98

<b>9 Analyzing Residuals</b>	<b>107</b>
9.1 Set up Your Model . . . . .	107
9.1.1 Steps to Evaluate Residuals . . . . .	107
9.1.2 Interpretation of Outputs . . . . .	108
9.1.3 Best Practice Summary . . . . .	108
9.2 Multivariate Linear Regression with Three Predictors Using Enter Term Selection . .	109
9.2.1 Results . . . . .	110
<b>10 Linear Regression Modeling with SPSS, Part 2: More about ANOVAs and Dummy Coding</b>	<b>115</b>
10.1 Overview . . . . .	115
10.2 Data . . . . .	115
10.3 Relationships Between Dummy Variables: Crosstabs and $\chi^2$ Tests . . . . .	115
10.3.1 Relationships with ELA Grades . . . . .	118
10.4 Using an ANOVA to Predict ELA Grades with Gender & IEP Status . . . . .	120
10.4.1 ANOVA Review . . . . .	120
10.4.2 Graphical Review of the Variables . . . . .	120
10.4.3 Using an ANOVA to Test Variables . . . . .	125
10.5 Linear Regression . . . . .	129
10.5.1 Creating an Interaction Term . . . . .	129
10.5.2 Computing a Linear Regression with an Interaction Term . . . . .	130
<b>11 Longitudinal Analyses: Why and How to Conduct Multilevel Linear Modeling</b>	<b>137</b>
11.1 Overview . . . . .	137
11.2 Comparison of Analyses of Longitudinal Data . . . . .	137
11.2.1 Explanations of Longitudinal Analyses . . . . .	137
11.3 Data Preparation and Manipulation . . . . .	145
11.4 Understanding the demographics.sav Data . . . . .	145
11.5 Understanding the discipline.sav Data . . . . .	146
11.6 Understanding the ef.sav Data . . . . .	146
11.7 Restructuring and Merging the Data Sets . . . . .	147
11.8 Wide-to-Long and Long-to-Wide Data Restructuring . . . . .	147
11.8.1 Wide-to-Long Data Transformation . . . . .	148
11.9 Long-to-Wide Data Restructuring . . . . .	148
11.10 One-to-Many Match Merge . . . . .	149
11.11 Conducting a Multilevel Model of Change . . . . .	150

11.11.1 Introduction to SPSS's Syntax . . . . .	150
11.11.2 Overview of Analyses . . . . .	152
11.11.3 Computing the Unconditional Means Model . . . . .	153
11.11.4 Interpreting the Syntax . . . . .	154
11.11.5 Interpreting the Results . . . . .	156
11.11.6 Computing the Unconditional Growth Model . . . . .	161
11.11.7 Interpreting the Syntax . . . . .	161
11.11.8 Interpreting the Results . . . . .	161
11.12 Univariate MLMs . . . . .	164
11.12.1 Gender . . . . .	165
11.12.2 Special Education . . . . .	171
11.12.3 Economic Distress . . . . .	175
11.12.4 Summary of Univariate Model Analyses . . . . .	177
11.13 Multivariate MLM . . . . .	178
11.13.1 Base Model . . . . .	178
11.13.2 Final Model . . . . .	181
11.14 Additional Resources & Topics . . . . .	184
11.14.1 Some Other Ways to Analyze Longitudinal Data . . . . .	184
<b>II Introduction to Psychometrics</b>	<b>187</b>
<b>12 NURS 925: Psychometrics Course</b>	<b>189</b>
12.1 Foundations of Modern Research Measurement . . . . .	189
12.2 Validity and Reliability . . . . .	189
12.3 Introduction to Factor Analysis & Exploratory Factor Analysis . . . . .	189
12.4 Confirmatory Factor Analysis . . . . .	190
<b>13 Exploratory Factor Analysis</b>	<b>191</b>
13.1 The Concept of Factor Analysis . . . . .	191
13.1.1 Role of Correlation/Covariance Matrix in Factor Analysis . . . . .	191
13.1.2 Factor Analysis Is Similar to the Linear Regression Analyses You Already (Should) Know . . . . .	192
13.2 Steps to Conducting an Exploratory Factor Analysis . . . . .	192
13.2.1 1. Estimate the Number of Factors . . . . .	192
13.2.2 2. Evaluate the Results . . . . .	196
13.2.3 3. Review Different "Rotations" of the Factors . . . . .	197
13.2.4 3. Interpretation . . . . .	198

<b>III</b>	<b>Guides to Using Software</b>	<b>201</b>
<b>14</b>	<b>Using Templates and a Reference Manager</b>	<b>203</b>
14.1	Overview . . . . .	203
14.2	Style Template . . . . .	203
14.2.1	Loading Templates in MS Word . . . . .	203
14.2.2	Loading Templates in LO Writer . . . . .	205
14.2.3	Using a Template . . . . .	205
14.3	Reference Manager . . . . .	212
14.3.1	RefWorks . . . . .	212
14.3.2	Zotero . . . . .	214
14.4	Additional Resources . . . . .	219
<b>15</b>	<b>Introduction to Excel</b>	<b>221</b>
15.1	Overview . . . . .	221
15.2	Filling in Series . . . . .	221
15.3	Pasting Special & Transposing . . . . .	222
15.4	Navigation . . . . .	222
15.4.1	Navigating & Selecting <i>Within</i> a Sheet . . . . .	222
15.4.2	Navigating <i>Between</i> Sheets . . . . .	223
15.5	Formulas . . . . .	223
15.5.1	General Steps for Entering Formulas (and Common Formulas to Use) . . . . .	223
15.5.2	if . . . . .	224
15.5.3	vlookup . . . . .	227
15.6	Pivot Tables & Charts . . . . .	227
15.6.1	Inserting/Creating a Pivot Table or Chart . . . . .	228
15.7	Basic Statistics . . . . .	228
15.7.1	Correlations . . . . .	228
15.7.2	<i>t</i> -Test . . . . .	229
15.7.3	ANOVAs . . . . .	230
<b>16</b>	<b>Introduction to Power and Sample Size Estimation Using Either G*Power or R</b>	<b>233</b>
16.1	The Relationship Between $\alpha$ , Power, Effect Size, and Sample Size . . . . .	233
16.2	Using G*Power or R to Estimate a Prior Sample Size Estimates . . . . .	235
16.3	G*Power . . . . .	235
16.3.1	Installing G*Power . . . . .	235

16.3.2 Citing G*Power . . . . .	235
16.3.3 Orientation to G*Power . . . . .	236
16.3.4 Estimating Required Sample Sizes . . . . .	237
16.4 R . . . . .	251
16.4.1 Comparing Two Independent Correlations . . . . .	252
16.4.2 Independent Samples <i>t</i> -Test . . . . .	252
16.4.3 Paired Samples <i>t</i> -Test . . . . .	252
16.4.4 Point-Biserial Correlation . . . . .	252
16.4.5 One-Way ANOVA . . . . .	253
16.4.6 Two-Way ANOVA . . . . .	253
16.4.7 Power Curves . . . . .	253
16.5 Additional Resources . . . . .	255
16.5.1 G*Power Guides & Tutorials . . . . .	255
16.5.2 Further Readings and Explanations . . . . .	255
16.5.3 Sample Size Estimations and Guidelines for More Complex Designs . . . . .	255
16.5.4 Online Tools . . . . .	256
<b>17 Introduction to SPSS &amp; Data Preparation</b>	<b>257</b>
17.1 Overview . . . . .	257
17.2 Orientation to SPSS . . . . .	257
17.2.1 Accessing SPSS Through Apporto . . . . .	257
17.2.2 Editing Global Options . . . . .	259
17.2.3 SPSS Windows . . . . .	260
17.3 Data Preparation & Cleaning . . . . .	265
17.3.1 Changing the Measurement Level of a Variable . . . . .	265
17.3.2 Creating a Variable Label for a Variable . . . . .	265
17.3.3 Setting Values Labels for a Variable . . . . .	266
17.3.4 Setting Missing Values . . . . .	268
17.3.5 Create Dummy Variables from Different Levels of a Nominal Variable . . . . .	268
17.3.6 Transform an Old Variable into a New One . . . . .	271
17.3.7 Standardize Variables . . . . .	274
17.4 Export a Table in APA Format to Word . . . . .	275
17.4.1 Steps to Export a Table in APA Format in SPSS: . . . . .	275
17.4.2 Example Export in Word: . . . . .	275
17.4.3 Summary . . . . .	275
17.5 Additional Resources . . . . .	276

<b>18 Data Exploration with R</b>	<b>277</b>
18.1 Common Exploration Commands . . . . .	277
18.2 Using SQL and tidyverse . . . . .	278
<b>References</b>	<b>281</b>
 <b>Appendices</b>	 <b>283</b>
<b>A Common Statistical Symbols</b>	<b>283</b>
 <b>B Common/Confusing Statistical &amp; Scientific Terms</b>	 <b>287</b>
B.1 Common/Confusing Statistical & Scientific Terms . . . . .	287
B.2 Terms for Different Types of Analyses . . . . .	292
 <b>C Statistical Analysis Decision Trees and Guides</b>	 <b>295</b>
C.1 References and Guides . . . . .	295
C.1.1 Correlations & Associations . . . . .	295
C.2 Simple Graphics . . . . .	296
C.3 Online Trees . . . . .	296

# List of Tables

6.1	Common Effect Size Measures of Mean Differences . . . . .	73
6.2	When to Use $\eta^2$ , $f$ , and $f^2$ . . . . .	76
6.3	Some Suggested Odds Ratios Corresponding to “Small,” “Medium,” and “Large” Effect Sizes Based on the Probability of the Event in the Reference Group (from Chen et al., 2010, p. 862) . . . . .	80
6.4	Effect Size Interpretations . . . . .	82
6.5	Formulas to Convert Between Common Effect Size Statistics . . . . .	83
6.6	Common Effect Size Statistics That Can Be Converted into Each Other . . . . .	84
6.7	Common Effect Size Statistics That <b>Cannot</b> Be Converted into Each Other . . . . .	84
10.1	Example Interpretation of Dummy-Coded Variables . . . . .	134
15.1	Key Commands to Navigate <i>Within</i> a Sheet in Excel . . . . .	222
15.2	Key Commands to Navigate <i>Between</i> Sheets in Excel . . . . .	223
15.3	Common Formulas in Excel . . . . .	224
15.7	vlookup Formula Match Modes in Excel . . . . .	227
15.9	$t$ -Test Tails in Excel . . . . .	229
15.10	$t$ -Test Types in Excel . . . . .	229
17.2	Example of Reverse Scoring Items . . . . .	271
A.1	Common Statistical Symbols . . . . .	283
B.1	Common/Confusing Statistical & Scientific Terms . . . . .	287
B.2	Terms for Different Types of Analyses . . . . .	292
C.1	Types of Correlation Statistics . . . . .	296



# Preface

This “book” is a companion to the statistics courses offered as part of the [Ph.D. in Nursing Research program](#) at [Hunter College, CUNY](#). It also provides supplemental materials for Hunter’s [Doctor of Nursing Practice program](#).

Very much a work in progress—and anyway intended to be in addition to and not instead of the many resources elsewhere—this companion currently contains:

- Lectures and essays on topics covered in the applied statistics curriculum:
  - NURS 60N lectures & materials (Chap. 1)
  - NURS 915 & 916 lectures & materials (Chap. 2)
  - A hands-on guide to writing Results sections (Chap. 3)
  - Discussions and activities expand upon the statistical concepts of significance (Chap. 4), correlations & partial/semipartial correlations (Chap. 5), effect size (Chap. 6), missing data (Chap. 7), and model building (Chaps. 8 – 11)
- Lectures and essays on topics covered in the introduction to psychometrics course, NURS 925 (Chaps. 12 & 13)
- Step-by-step guides to using various stat-related software (Chaps. 14 – 18)
- Tables of common statistical abbreviations and terms along with their meanings (Apps. A & B)
- A small collection of decision trees (flowcharts) to decide which analysis to conduct (App. C)

## Colophon



This companion is created 2022 – present by William Ellery Samuels, Ph.D., under a [Creative Commons Attribution–NonCommercial–ShareAlike license](#). (Note that the “Published” date given above is in fact the date of the most-recent revisions.)

Light-themed HTML-version typefaces are [Bona Nova](#) for text, [Ubuntu Mono](#) for code blocks, and [Latin Modern Math](#) for formulas. The dark HTML is theme is unmodified [solar](#), which is based on [Ethan Schoonover](#)’s eye-friendly [Solarized](#) dark theme.

PDF-version typefaces are [TeX Gyre Bonum](#) for both text & formulas and TeX Gyre [Adventor](#) for code blocks.

OpenAI’s [ChatGPT](#) has been occasionally used for clarity and content review—and extensively to debug code errors and issues (I’m looking at you, [Quarto](#)).

This *Companion* is produced through [Bookdown](#) in the [RStudio](#) environment with:

```
sessioninfo::session_info()
```

```
- Session info -----
```

```
setting value
version R version 4.5.1 (2025-06-13)
os      Ubuntu 24.04.3 LTS
system  x86_64, linux-gnu
ui      X11
language (EN)
collate en_US.UTF-8
ctype   en_US.UTF-8
tz      America/New_York
date    2025-09-25
pandoc  3.1.3 @ /usr/bin/ (via rmarkdown)
```

```
- Packages -----
```

```
package * version date (UTC) lib source
cli      3.6.2  2023-12-11 [3] CRAN (R 4.3.2)
digest   0.6.34  2024-01-11 [3] CRAN (R 4.3.2)
evaluate  0.23     2023-11-01 [3] CRAN (R 4.3.2)
fastmap   1.1.1    2023-02-24 [3] CRAN (R 4.2.2)
htmltools 0.5.7    2023-11-03 [3] CRAN (R 4.3.2)
jsonlite  1.8.8    2023-12-04 [3] CRAN (R 4.3.2)
knitr     1.45     2023-10-30 [3] CRAN (R 4.3.2)
rlang     1.1.3    2024-01-10 [3] CRAN (R 4.3.2)
rmarkdown 2.25     2023-09-18 [3] CRAN (R 4.3.2)
sessioninfo 1.2.2    2021-12-06 [3] CRAN (R 4.1.2)
xfun      0.41     2023-11-01 [3] CRAN (R 4.3.2)
```

```
[1] /home/wes/R/x86_64-pc-linux-gnu-library/4.5
```

```
[2] /usr/local/lib/R/site-library
```

```
[3] /usr/lib/R/site-library
```

```
[4] /usr/lib/R/library
```

```
-----
```

© — —  $\oint$  =====  $\gamma$

## **Part I**

# **Applied Statistics Curriculum**



The content of the sequence of stat courses is:

## **NURS 6oN**

1. Major Principles
  - a. Randomness & Variables
  - b. Samples vs. Populations
    1. Evaluation vs. Research
  - c. Descriptive vs. Inferential Statistics
    1. Parametric vs. Non-Parametric Analyses
  - d. Related to it is this [online Demonstration]
2. Frequencies & Counts
  - a. Frequencies & relative frequencies
  - b. Probabilities
    1. Risks & risk ratios
    2. Hazards & hazard ratios
  - c. Odds & odds ratios
  - d. Contingency / cross tables
    1. Fisher's exact test
  - e. Important distributions
    1. Normal distribution
    2.  $\chi^2$  distribution
3. Measuring & Testing Differences
  - a. Review of Assumptions in Inferential Statistics
  - b. Hypothesis Testing
  - c. Signal-to-Noise Ratio
  - d. Common Tests:  $t$  &  $F$
4. Power and Effect Size
  - a. Review & Elaboration of Hypothesis Testing
  - b. Power
  - c. Effect Size
5. Association & Causation
  - a. Individual Differences and Correlations
  - b. Types of Correlation Statistics
  - c. Partial and Semipartial Correlations
  - d. Concerning Causality
6. The ANOVA Family of Tests
  - a. Basic Concepts of ANOVAs
  - b. Main Effects & Interactions
    1.  $R^2$  &  $\eta^2$
  - c. Reading Source Tables
  - d. Types of ANOVAs
  - e. Family-Wise Error, Post hoc, & Planned Comparisons

**NURS 915 & 916**

1. Overview & Review
2. Handling Data
3. Presenting Data
4. Power & Significance
  - a. Post hoc power
  - b. Sample size estimation
5. Introduction to Linear Regressions
  - a. Method of Ordinary Least Squares
  - b. Model Assumptions
6. Ordinary Least Squares & Maximum Likelihood Estimation
  - a. General & Generalized Linear Models
7. Tests of Model Fit
  - a. Information Criteria
  - b. Residual Analysis
  - c. Stepwise Analysis
  - d. Bootstrapping
  - e. Missing Values & Outliers
8. Occurrence, Association, & Causation
  - a. Counterfactuals & Hill's Criteria
  - b. Confounds, Mediators, & Moderators
9. Logistic Regression
  - a. Multinomial & Ordinal Logistic Regression
10. Hierarchical Regression
11. Longitudinal Analyses
  1. Pre-Post Differences ("Differences in Differences")
  2. (Repeated-Measures) ANCOVAs with Pretest as Covariate
  3. Multilevel Models of Change
  4. Interrupted Time Series Analysis
12. Robust Statistics
  - a. Bootstrapping
  - b. Missing Values & Outliers
13. Structural Equation Modeling

**NURS 925**

1. Foundations of Measurement and Scaling
  1. Psychophysics & Psychometrics
2. Validity
  1. Traditional, Trinity View
  2. The 1999–2014 *Standards* & Validity as "Use"

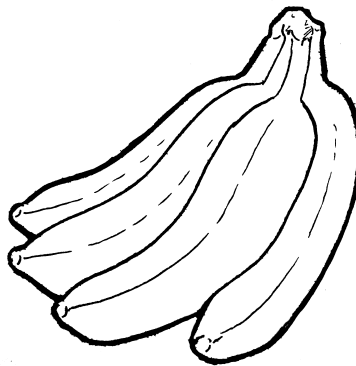
### 3. Reliability

1. Classical Measurement Theory View
2. As a Measure of a Unitary Construct
3. As a Measure of Internal Consistency
  1. Cronbach's  $\alpha$
  2. Kuder-Richardson Formulae 20 & 21
4. Other Forms (Test-Retest, etc.)

### 4. Factor Analysis

1. Concept and Basic Ideas
2. Eigenvalues
3. Exploratory Factor Analysis
  1. Uses and abuses
4. Confirmatory Factor Analysis
  1. Measures of Model Fit

### 5. Return to Structural Equation Modelling





# Chapter 1

## NURS 6oN: Foundations of Biostatistics for Nursing Research and Evidence-Based Practice

This “chapter” contains links to the presentations and materials covered in the first course of this curriculum, NURS 6oN.

### 1.1 Major Principles

- Randomness & Variables
- Samples vs. Populations
  - Evaluation vs. Research
- Descriptive vs. Inferential Statistics
  - Parametric vs. Non-Parametric Analyses

Related is this [online demonstration of the Central Limit Theorem](#)

### 1.2 Frequencies & Counts

- Frequencies & Relative Frequencies
- Probabilities
  - Risks & Risk Ratios
  - Hazards & Hazard Ratios
- Odds & Odds Ratios
- Contingency / cross tables
  - Fisher’s Exact Test
- Important Distributions

- Normal Distribution
  - \* Q-Q Plots
- $\chi^2$  Distribution

Related is this [Shiny app](#) to practice matching the distribution to the Q-Q plot

### 1.3 Measuring & Testing Differences

- Review of Assumptions in Inferential Statistics
- Hypothesis Testing
- Signal-to-Noise Ratio
- Common Tests:  $t$  &  $F$

### 1.4 Power and Effect Size

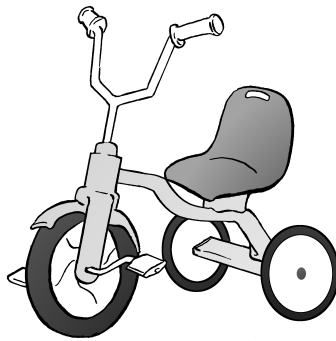
- Review & Elaboration of Hypothesis Testing
- Power
- Effect Size

### 1.5 Association & Causation

- Individual Differences and Correlations
- Types of Correlation Statistics
- Partial and Semipartial Correlations
- Concerning Causality

### 1.6 The ANOVA Family of Tests

- Basic Concepts of ANOVAs
- Example of Main Effects & Interactions
- Types of ANOVAs
- $t$ - &  $F$ -Tests Roles & Nature
- Post hoc Comparisons
- Reference Tables for ANOVA Terms





## Chapter 2

# NURS 915 & 916: Applied Statistics 1 & 2

This “chapter” contains links to the presentations and materials covered in the Applied Statistics courses of the curriculum, NURS 915 and 916.

### **2.1 Review of Some Major Principles & Practices in Biostatistics**

- Variability & Randomness
- Levels of Measurement
- Descriptive & Inferential Statistics
- Sources of Variance and the Signal-to-Noise Ratio
- Designing and Answering Questions
- Hypothesis Testing

[Recording of Lecture](#)

### **2.2 Presenting Results**

- Visualization
  - Self Sufficiency
  - Efficient Information Transfer
  - Data-to-Ink Ratio
  - Follow Conventions & Readers' Expectations

#### **2.2.1 Writing Results**

- Writing Results Sections
  - Tell a Story

- Use Figures & Tables as Talking Points
- Use Statistics as Citations to Support Assertions

Additional information and resources are given in Chapter 3.

## 2.3 Introduction to Linear Models

- Review of Underlying Concepts in Inferential Statistics
  - Additional information and resources are given in Chapter 5
- Understanding Linear Models
  - Basic Concepts
  - Linear Models Vs. Correlations
  - Understanding the Linear Equation
  - Linear Models Vs. ANOVAs
- Terms in Linear Models
  - Adding Terms to Models
  - Signal-to-Noise in Linear Models
  - Generalized Linear Models
- An Example
- Further Considerations
  - Multicollinearity
  - Independence of Cases

[Recording of Lecture](#) - A Zoom recording from a previous semester

[PDF of Lecture](#)

[DOCX of Lecture](#)

Additional information and resources—including steps to conducting them in SPSS—are given in Chapters 8 and 10

## 2.4 Testing Models Theoretically

- Review of Linear Regression Model
- Partialling out Variance
- Combining Similar Sources of Variance
- Ostensible & Non-Ostensible Variables
- Model Fit

[Recording 1](#) - A Zoom recording from a previous semester, this recording contains a review of linear models and introduction to tests of model fits.

[Recording 2](#) - An other Zoom recording from a prior semester, this covers an explanation of ANOVAs and their qualities vs. general linear models

## 2.5 Analyses of Longitudinal Data

- Longitudinal analyses, including some of their benefits and challenges
- A brief comparison of the merits of pre-post difference scores, including pretest covariates in ANCOVAs, and repeated-measures ANOVAs.
- An introduction to the sorts of multilevel models of change that Singer & Willett (2003) describe

### Recording

Additional information and resources—including steps to conducting them in SPSS—are all currently located in Chapter 11.

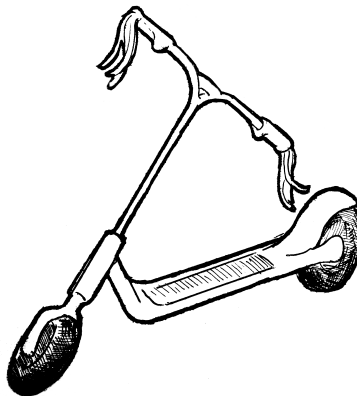
## 2.6 Logistic Regression

- Logistic regression vs. general linear regression
- Explanation of the math
- Testing effects & model fit
- Types of logistic regression
- Examples

## 2.7 Structural Equation Modeling

Structural equation models can be considered a sort of bridge between generalized linear regression and the factor analyses we'll cover in NURS 925, Psychometrics.

- Core Concepts
- Mechanics of SEMs
- Comparing Models
- Example of SEMs





## Chapter 3

# Writing Results Sections

*Just to try make it sound like you wrote it that way on purpose. – The French Dispatch*

### 3.1 Overview

This chapter is in some ways an abbreviated version of *Section 8.3: Introduction to Linear Regression Models* in Chapter 8. This sheet is intended to orient one subset of those analyses toward working within Word to write results and less on explaining how to use SPSS to prepare and analyze data. We will thus:

1. Review a few options in SPSS that will help prepare tables and figures for addition to a Results section created in Word.
2. Walk through a set of analyses that—in miniature—emulate the sorts of exploratory analyses one might do with the data on executive functioning slopes and adolescent academic performance & disciplinary actions.
3. Add a few tables and figures from those analyses to a Word document that uses an APA-formatted template
4. Write pieces of a Results section around those tables and figures

### 3.2 Setting Global Options in SPSS and Using a Word Template

#### 3.2.1 Setting Global Options in SPSS

The following two action simply change the fonts for the figures and tables we'll create. This is, of course, a small change, but most manuscripts use a serif font—and often the vaunted<sup>1</sup> Times New Roman font. Now, all of your figures and tables will use that font.

##### 3.2.1.1 Charts

1. In the SPSS Edit > Options menu, click on the Charts tab in the dialogue that appears

---

<sup>1</sup>Worn out, if you ask me

2. Change the Style cycle preference from Cycle through color only to Cycle through patterns only
3. Now, change the Font from SansSerif to Times New Roman.

### 3.2.1.2 Tables

1. Under Edit > Options menu, now click on the Pivot Tables tab in the dialogue that appears
2. Under TableLook, select APA\_TimesRoma\_12pt
3. Not all table will now be formatted this way, but the pivot tables and a few more will

### 3.2.2 Using a Word Template

Using templates takes some getting used to, but it can be both time saved in the long run and even required by publishers like Elsevier, SAGE, and Springer.

To use a template in Word:

1. Download this Word template: [APA\\_7th\\_Ed\\_Template.dotx](#)
  - Note that the extension for that file is .dotx—not .docx. That “t” denotes that this is a template<sup>2</sup>
2. After saving that file to a more useful location than the downloads folder (ahem), open it in Word.
  - Note that I’ve added some language into the template to help you both know most of the elements in it and to serve as a structure for manuscripts. You can remove any and all of the text in the template (or, of course, replace it with your own); the “magic” of a template isn’t in the textual content. Indeed, that’s the *point* of a template: to separate the actual content that’s written from the styling of that content. The
3. Choose to Save as a **.docx** file—not a .dotx file—perhaps calling it something like “Ng16\_Results\_Exercise.docx.” This will preserve the template and change the extension to what is now the correct format since we will indeed be adding content and making this no longer a template to be used for other documents.

I have given a fuller explanation of how to use Word templates in Chapter 14, so I will not cover the details of accessing and updating styles and sections in this guide.

## 3.3 Data Prep and Cleaning

The data that we will be using in the EF\_Slope\_Data.sav file are pretty clean and ready, but there are nonetheless a few tasks we will perform on them. Although these are not real data, they are loosely based on real data, in any case, please treat them as if they were real.

---

<sup>2</sup>Does this mean that the “c” in .docx denotes a normal document? Of course not, this is Microsoft; they don’t try to be systematic in what they do. As an other aside, the “x” at the end was added by Microsoft because they were forced to. The world of computing once again moved beyond Microsoft’s kludgy programming with growing support for Open Office’s “extensible markup language” (XML), and Microsoft had to fundamentally change how it generated Word files to catch up again with advances in technology. And the advance in technology they had to catch up with that time was simply what we’re doing right now: Separate out the content from the styling. What you say from how it looks. So you can focus on saying it knowing it will look good. To do this, Microsoft had to totally revise the way it made Word (and other types of) files.

Load those data, either with the initial dialogue box that opens by default when SPSS loads or by choosing File > Open > Data... from the menu bar and selecting EF\_Slope\_Data.sav wherever you have it stored.

### 3.3.1 Adding Variable Value Labels

1. Most of the dummy-coded data have their values labeled, but DBT does not. After shaking your head at the forgetfulness of your instructor, open the Variable View tab of the Data Editor window.
2. Click on the ellipsis button that appears when you click on the cell that is intersected by the DBT row and Values column.
3. In the Value Labels dialogue box that opens, enter a 0 (a zero) in the Value: field and Did Not Participate in the Label: field; then click on the Add button.
4. Now, enter a 1 in the Value: field and Participated in the Label: field before clicking Add and then OK.

### 3.3.2 Changing Measurement Scale

1. When looking in the Variable View tab in the Data Editor window, we see that that same DBT variable is listed in the Measure field as a Scale variable, indicating that SPSS is treating it like a real number. A zero in a dummy variable indicates that that trait is not present (that the student did not participate in the DBT program, the student is not female, does not experience excessive economic hardship, etc.); the zero is not a zero on a continuous scale.
2. We can change this easily enough in the Variable View tab of the Data Editor window. In that tab, click on the Measure cell in the DBT row. From the drop-down menu, choose to change it to Nominal.
3. Note that the Type will remain Numeric. Changing the Measure to Nominal lets SPSS know that this number simply indicates the presence of absence of the trait denoted by the dummy variable.

### 3.3.3 Creating Standardized Variables

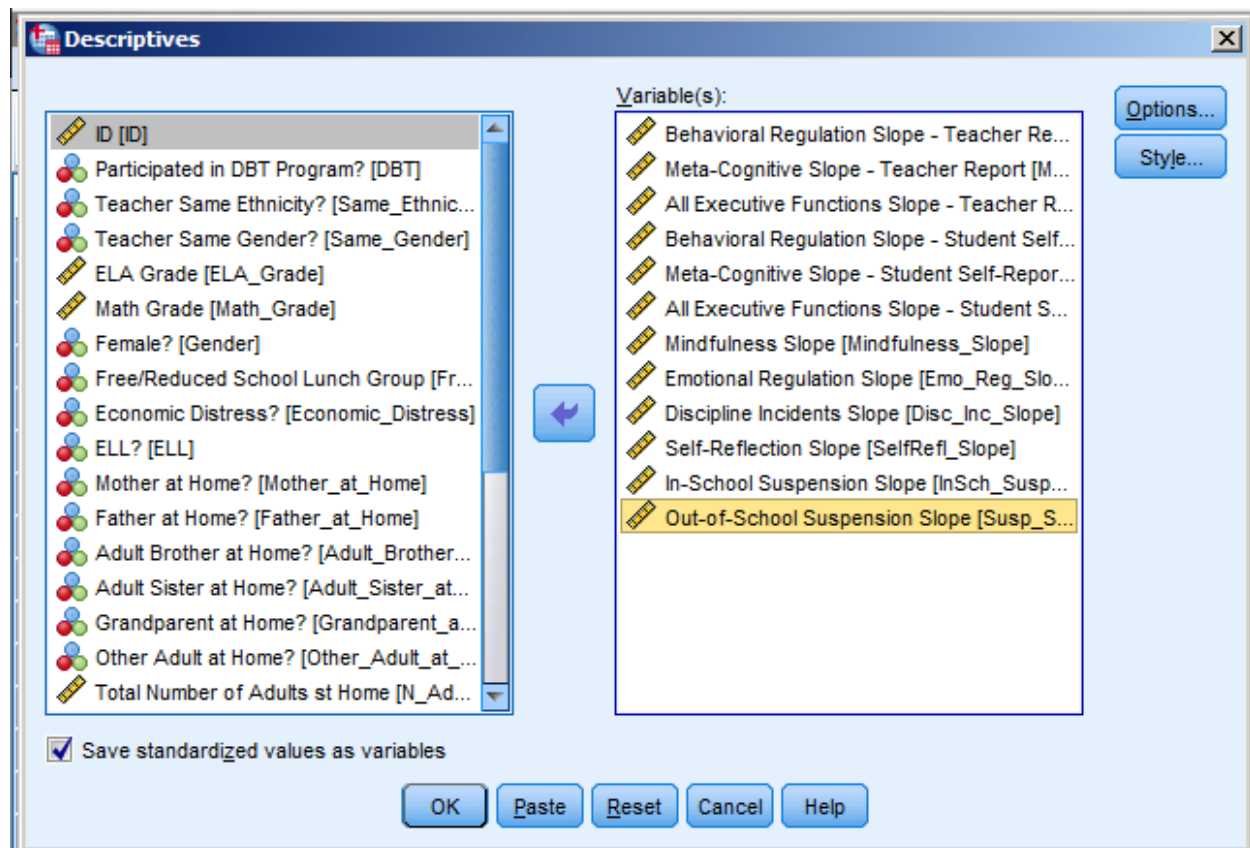
I computed the slopes from non-standardized scores. The mean for these slopes is nearly zero (e.g., -0.0018 for Beh\_Reg\_I\_Slope), but if we convert them to truly standardized scores, we would occur these advantages:

- Being z-scores, you can get a sense of significant values (and differences): a z-score of ~1.65 is significantly different than zero in a one-tailed test where  $\alpha = .05$  (and two z-scores that differ by ~1.65 are significantly different).
  - z-Scores of ~1.96 are significantly different at  $\alpha = .05$  in a two-tailed test. Smaller z-score values will be significant when the degrees of freedom for formal tests are larger.
  - Comparisons between standardized variables can be made directly since they are on the same scale. This is true not only within your data set, but also—to some extent—between sets of data. (This is contingent on the assumption that any measurement biases are similar between the two sets of data.)
  - You can remove the intercept term from models where all variables are standardized. This frees up a degree of freedom and makes for a somewhat easier interpretation of the model terms.

- \* Sometimes (actually, often) predictors are correlated with the intercept. Removing the intercept by standardizing our variables lets us “remove” the effect of the intercept and instead let that variance be placed in with the variables of interest. This can clarify our analyses, especially by usually reducing the chance of a Type 2 error.
  - Similarly, if some of the predictors are strongly inter-correlated (there is strong “collinearity” as it’s called), standardizing and removing the intercept may help reduce that.
- The model weights for each variable will be expressed in the same scale as a correlation; if we square these standardized weights, we obtained the proportion of variance accounted for in the criterion by that particular predictor (more on this later; I just wanted to add it to the list here).

Generating standardized scores in SPSS is a bit counter-intuitive. To create them:

1. Click on Analyze > Descriptive Statistics > Descriptives
2. In the dialogue box that opens, select the variables you would like to standardize. Let’s select all of the variables that are slopes.
3. In that same dialogue box, select Save standardized values as variables under the list of variables to choose from which to add to the Variable(s) pane:



4. This will generate descriptive statistics in the Output window, but our interest right now is the changes made to the data set itself: At the far right of the matrix—just past Disc\_Inc\_Y4, is a set of new variables. You will see that these are the slope variables, each now prepended with a Z to indicate that these are z-scores, i.e., standardized.

5. (Note that you can also compute standardized scores in the Transform > Compute Variables dialogue. This interface is a lot more flexible, but doesn't have a simply function to standardize.)
6. We'll be using ZAll\_EFs\_SR\_Slope more, so let's move up higher in the list for easier access. To do this, go the Variable View tab in the Data Editor, and scroll down to the set of standardized variables. Single left-click on ZAll\_EFs\_SR\_Slope; holding the mouse button down, drag that variable up, e.g., to right before the non-standardized set, right below Same\_Gender.
7. (We can move more than one variable this way. We could, e.g., left on the first standardized variable (ZBeh\_Reg\_T\_Slope), then hold down the Shift key and single left-click on the last one (ZSusp\_Slope) to select them all. Then single left-click on this highlighted set; keeping the mouse button down, drag this list up.
8. Since we're also using Economic Distress more, we could left-click on ZAll\_EFs\_SR\_Slope, hold down the Ctrl button, left-click on Economic Distress, and then hold the mouse button down to drag them both up. Finally, note as well that we could have done this in the Data View tab, but I think it's easier in the Variable View.)

### 3.3.4 Creating Dummy Variables

Recall that a **dummy variable** is a dichotomous variable that indicates whether something is present. For example, Economic Distress here lets 1 denote the presence of economic distress (beyond a pre-established threshold) and lets 0 denote the absence of that much economic distress. Similarly, Gender here lets 1 denote the "presence" of a female and 0 denote "not female."

Some linear regression models **require** that nominal variables be represented by dummy variables. Even when a model doesn't require it—like an ANOVA—adding variables as dummy variables can help with interpreting the differences between nominal categories (since we don't have to then conduct *post hoc* analyses after the ANOVA to see where the differences were).

For most of the relevant variables here, I've already transformed them into dummy variables. Again, though, there is one I didn't: Ethnicity. We'll do that now.

1. There are a **few** ways to create dummy variables in SPSS; we'll use the method perhaps easiest for when we are creating a good number of dummy variables to accommodate several categories.
2. In SPSS, choose Transform > Create Dummy Variables
3. Under Dummy Variable Labels, select Use values. This will use what is given in the cells of the Ethnicity variable to create dummy variable names. If we had created variable labels (like we did, e.g., for DBT), then we could consider using those labels instead.
4. In the Variables list that appears in the dialogue box that opens, choose Ethnicity to place in the Create Dummy Variables for: field.
5. Under Main Effect Dummy Variables, leave Create main-effect dummies selected.
6. Enter a Root Names (One Per Selected Variable):. These will all be ethnicities, of course, so ethnicity is a logical choice here<sup>3</sup>. (Note that we cannot make the root name the exact same as the name of another variable; for example, here we couldn't have used Ethnicity as the root name.)
7. We do not have enough levels to create 2- or 3-way effects. If we did, choosing this would allow us to combine levels to look at more complicated relationships between them.

---

<sup>3</sup>One could argue that we should be using "race/ethnicity" throughout.

8. The dummy variables should now appear at the right right side of the data matrix.
9. As you can see when looking at these new variables, when SPSS creates dummy variables, it appends the root name (here ethnicity) with a number, creating ethnicity\_1, ethnicity\_2, etc. without any other explanation in the data about what level is indicated. Therefore, when creating more than one dummy variable (like we did here), it is important to look at SPSS's output; this shows us what level is being indicated in each variable. I think it's useful to change the variable names to make it clearer what is being indicated, so we'll do that after one more step.
10. Normally when creating dummy variables, one must first decide what level is the "reference level," the level against which all other levels will be compared. In some cases, this choice is easy: For the DBT dummy variable, the control group is the clear choice for the reference level; when we set that as the control, then a significant DBT variable effect in a model means there is a significant effect of participating in the DBT program. This means that there is typically one less dummy variable than there are levels to the original variable.

For example, with ethnicity, we could set one ethnic category to be the reference—the one against which all others are compared. There's no *a priori* reason to choose any particular ethnic group here; going alphabetically, if we made American Indian the reference group for all other levels, then we would not need a separate variable for American Indian: If a person had 0s for all of the dummy variables, that would mean that that person was American Indian. How SPSS creates dummy variables in this method, though, there was no particular group chosen to be the reference group; there is a dummy variable "yes / no" for each level—including for NA, the data were missing.

11. The variable Labels have been made informative by SPSS, so we can see which is which. Nonetheless, the variable names are not informative. We can use this peccadillo of SPSS to take a moment to clean up these dummy variables.
12. To go quickly from the Data View to the right place in the Variable View, double-click on ethnicity\_1's column heading in Data View. This will take you right to that variable in the Variable View.
13. Since there are so few American Indians<sup>4</sup>, we will not be analyzing them, and so we can right-click on whichever of the dummy variables we just created has Ethnicity=American\_Indian as its Label in the Variable View. This may be the first dummy variable for you (i.e., ethnicity\_1) or another one, perhaps the second (i.e., ethnicity\_2).

If you cannot right-click, then you can main-click on row number for the American Indian dummy variable, and then either hit the Delete key or select Edit > Clear. This will delete that variable.

14. There are arguably enough Asian-Americans, and certainly enough African- and Hispanic-Americans, so let's not delete those three. Instead, change the label of the variable whose Label is Ethnicity=Asian to, e.g., ethnicity\_asian. and the others to, e.g., ethnicity\_african and ethnicity\_latin.
15. Still in Variable View, right-click on the row header for whichever variable's Label is Ethnicity=Multiracial and choose Descriptive Statistics (or click on Analyze > Descriptives > Descriptive Statistics...). There are only 5 (7%) students who reported being multiracial; in "real" analyses, I would keep them, but to keep things tidier here, let's delete this variable, too; with that row still selected (from looking at the descriptives), hit the Delete button to delete it.

---

<sup>4</sup>Feh, left in the world after what we've done to them, let alone here

16. We can delete whichever dummy variable has nothing after Ethnicity= in its Label; these are missing values since there are many better ways to evaluate missing data. There are enough Whites to change the name of the variable with the Ethnicity=White label to, e.g., ethnicity\_european.

We have now created a nice set of dummy variables. Again, this set does not exhaustively explain the ethnicities of the students. Instead, we have created one that will allow us only to look at the effects of ethnic categories that we expect to be possibly informative.

One final word about these dummy variables. How the school records ethnicity, a student can either be Hispanic or Black, White, Asian, etc. Except perhaps for “Multiracial,” a student could not be both Hispanic and, e.g., Black. However, with these dummy variables—if we knew—we could in fact have a student be Hispanic (ethnicity\_latin), Black (ethnicity\_african), White (ethnicity\_european), and Asian (ethnicity\_asian) simply by having a 1 in each of those variables (or by creating 2- or 3-way variables in SPSS, but—frankly—I rarely see the value of that over just having a dummy variable for each).

## 3.4 Exploring the Data

### 3.4.1 Descriptives

1. Go to Analyze > Descriptive Statistics > Descriptives
2. To the Variables field, add Spec\_Ed, DBT, Economic\_Distress, All\_EFs\_SR\_Slope, and Disc\_Inc\_Slope. Note that—like Microsoft—SPSS tends to use a lot of tiny little windows for no good reason. You can expand them, but you have to expand *each* of them. It may be easier to only look at variable names, not also/instead the variable labels. You can do this by right-clicking on one of the variables in the list and selecting Display Variable Names.
3. The Options are fine, but you can add (a few) more stats. Note that Distribution stats are nearly always under-informative at best, misleading at worst

Including just the variable noted above should create a table that looks like this:

<i>Descriptive Statistics</i>					
	N	Minimum	Maximum	Mean	Std. Deviation
Special Education Status	507	0	1	.40	.490
Participated in DBT Program?	670	0	1	.19	.389
Economic Distress?	578	0	1	.62	.485
All Executive Functions Slope - Student Self-Report	326	-.50000000	.642857143	.006745607	.083688175
Discipline Incidents Slope	357	-1.5000000	1.50000000	-.04247114	.647913097
Valid N (listwise)	152				

We can prettify the table a bit<sup>5</sup>:

1. Double-clicking into the table so that we may edit it
2. Left-click to select the cells that have a whole bunch of decimals places (i.e., The Minimum through Std. Deviation columns in the All Executive Functions Slope - Student Self-Report and the Discipline Incidents Slope rows)
3. Right-click and select Cell Properties
4. Under the Format Value tab, change the Decimals (near the bottom) to 2, and click OK

The table should now look like this:

*Descriptive Statistics*

	N	Minimum	Maximum	Mean	Std. Deviation
Special Education Status	507	0	1	.40	.490
Participated in DBT Program?	670	0	1	.19	.389
Economic Distress?	578	0	1	.62	.485
All Executive Functions Slope - Student Self-Report	326	-.50	.64	.01	.08
Discipline Incidents Slope	357	-1.50	1.50	-.04	.65
Valid N (listwise)	152				

When the table is generated, we can simply right-click on it and select Copy before then pasting it into Word.

#### **i** Note

Older convention—and some current journals—have tables and figures relegated to the end of a manuscript. This was done because publishers used to actually take a photograph of these elements and include that photo in the published article. Clearly, this is no longer done, and thus this is no longer needed. APA 7th (p. 43) says that “[t]ables and figures may be embedded within the text after they have been mentioned, or each table and figure can be displayed on a separate page” at the end of the manuscript. I have never met anyone who prefers to read a manuscript while flipping back and forth to something at the end, so I recommend indeed placing them within the Results section<sup>6</sup>.

Once it's in Word, add a title to this table, e.g., *Descriptions of Demographic and Outcome Variables*. To do this, please follow the steps in the *Adding and Captioning a Figure or Table* section of the *Using Templates and a Reference Manager with MS Word* chapter.

<sup>5</sup>Well, you can do a lot both within SPSS and Word to make tables look quite good. I'm now simply showing one way that makes a relatively significant change. Nonetheless, you may want to explore the other options presented when you go through the following steps.

<sup>6</sup>Note that some journals do still require them to be at the end.

### 3.4.2 Guided Data Exploration

SPSS's Explore function can be rather useful. Part of it replicates what we did using Descriptives, but it adds several other ways to view the data and relationships within them.

1. Go to Analyze > Descriptive Statistics > Explore
2. To Dependent List, add All\_EFs\_SR\_Slope and Disc\_Inc\_Slope. These are, of course, the continuous variables from the subgroup we've been using. We don't have to use only continuous variables in the Dependent List, however.
3. To the Factor List, add Spec\_Ed, DBT, and Economic\_Distres
4. Under Plots, leave Boxplots to Factor levels together and perhaps choose both Stem-and-leaf and Histogram
  - I don't see the value here in selecting Normality plots and tests, but it could be useful—if overly sensitive—in other instances
5. Under Options select Exclude cases pairwise, remember that excluding listwise is really never advisable. Report values would also be useful, except that here we have so many it would overwhelm us.
6. Under Display in the main dialogue, make sure Both is selected. SPSS is good for exploring data since it can easily generate lots of output.

Note that you might get error messages, mostly related to converting “XML characters.” This is simply from using the data across operating systems with different conventions for reading characters<sup>7</sup>. It is certainly worth reading any error messages you get—and you will get them from time to time. Here, they are not important, though.

#### 3.4.2.1 Stem-and-Leaf Plots

A stem-and-leaf plots is frequency distribution, but one that also gives information about the values in each stack or “stem.” Stem-and-leaf plots were more commonly used before the easy graphics of computers. They are still quite useful, though, displaying a good amount of information efficiently.

The one detailing the All Executive Functions Slope - Student Self-Report scores for students with No Diagnosed Disability looks like this:

---

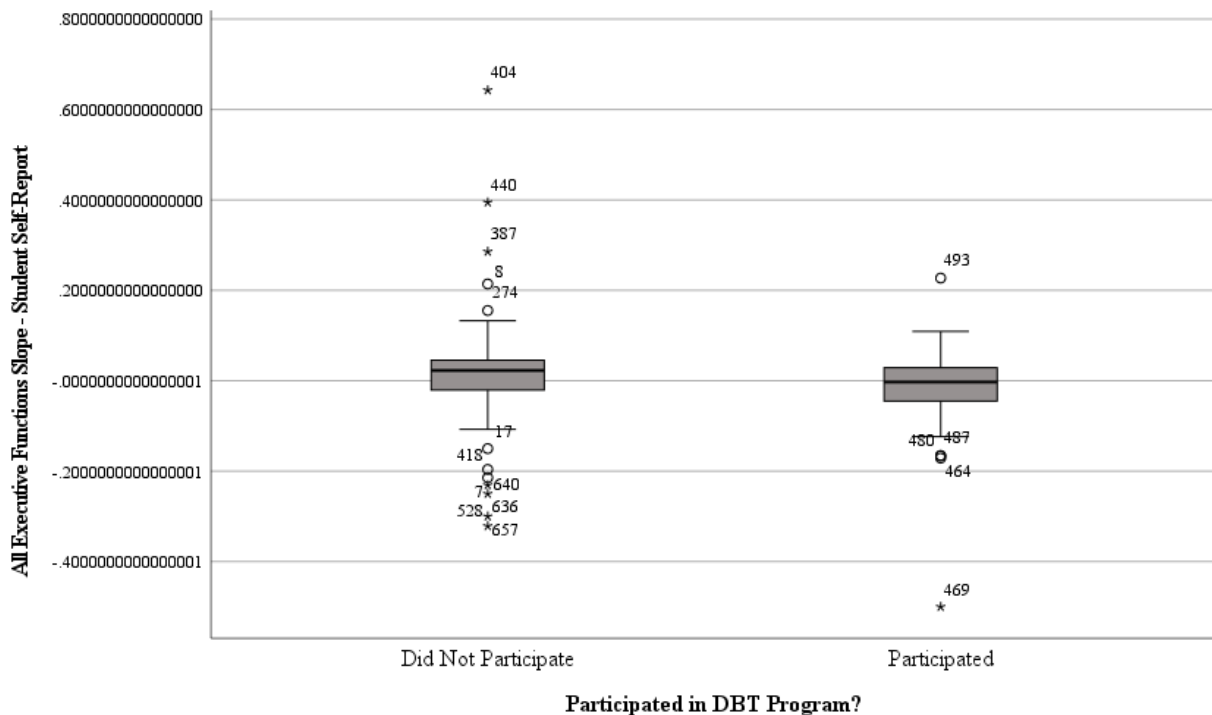
<sup>7</sup>And yes, this is a rare moment when I didn't take the obvious chance to bash Microsoft for their lack of compliance to established conventions.



### 3.4.2.2 Box-and-Whisker Plots

Like stem-and-leaf plots, [box-and-whisker plots](#) summarize the distribution of scores without making any assumptions about them. They simply describe the distribution. Box-and-whisker plots provide a bit less information than stem-and-leaf plots, but they may be a bit easier to read at a glance.

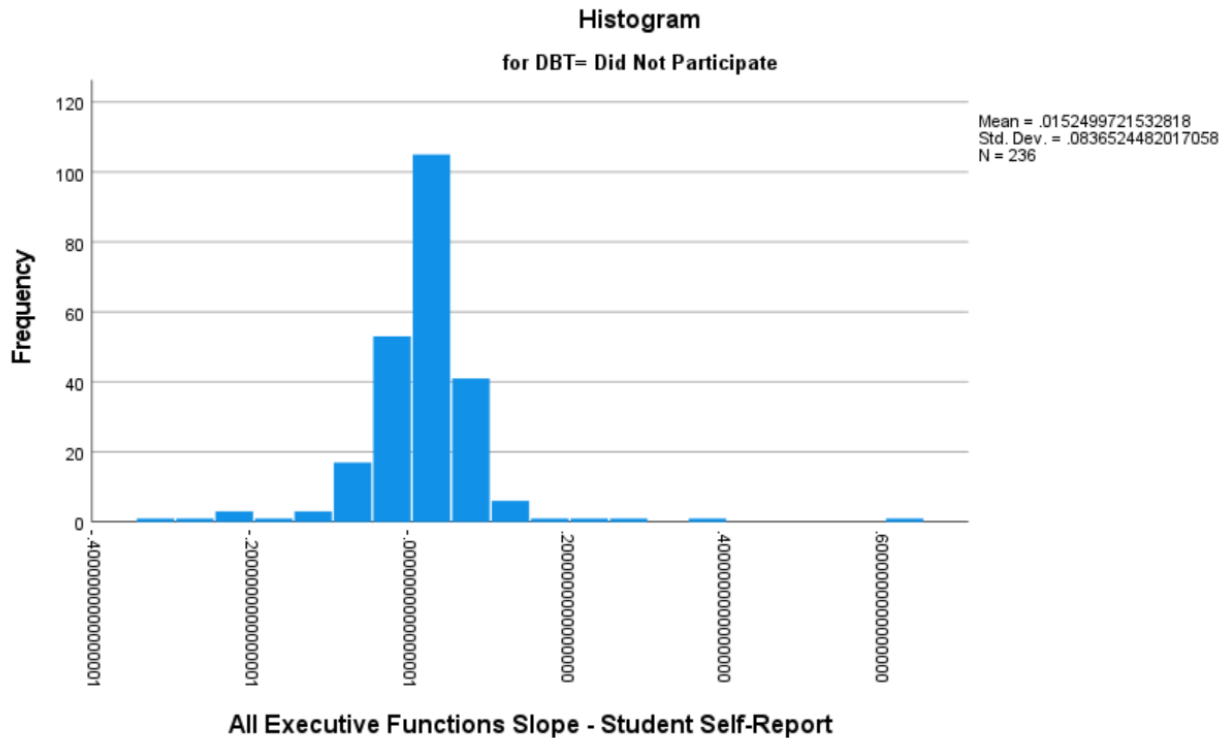
The one for student self-reported GEC slopes by DBT participation is:



Box-and-whisker plots are easily confused with figures that have error bars on them, but these box-and-whisker plots are different. The thick, black line in the middle of the grey box indicates the **median**. The grey box around that thick line indicates the range of scores that contains the middle half of the data; in other words, the space between the top of that grey box and that thick line contains the 25% of the scores above the median, and the space between the bottom of that grey box and the thick line contains the 25% of the scores just below the median.

The “whiskers” are those thin “Ts” that extend out above and below the grey box; these contain (just about) the upper- and lower-most 25% of the data. I say “just about” since scores that could be considered outliers are shown individually beyond the ends of those whiskers.

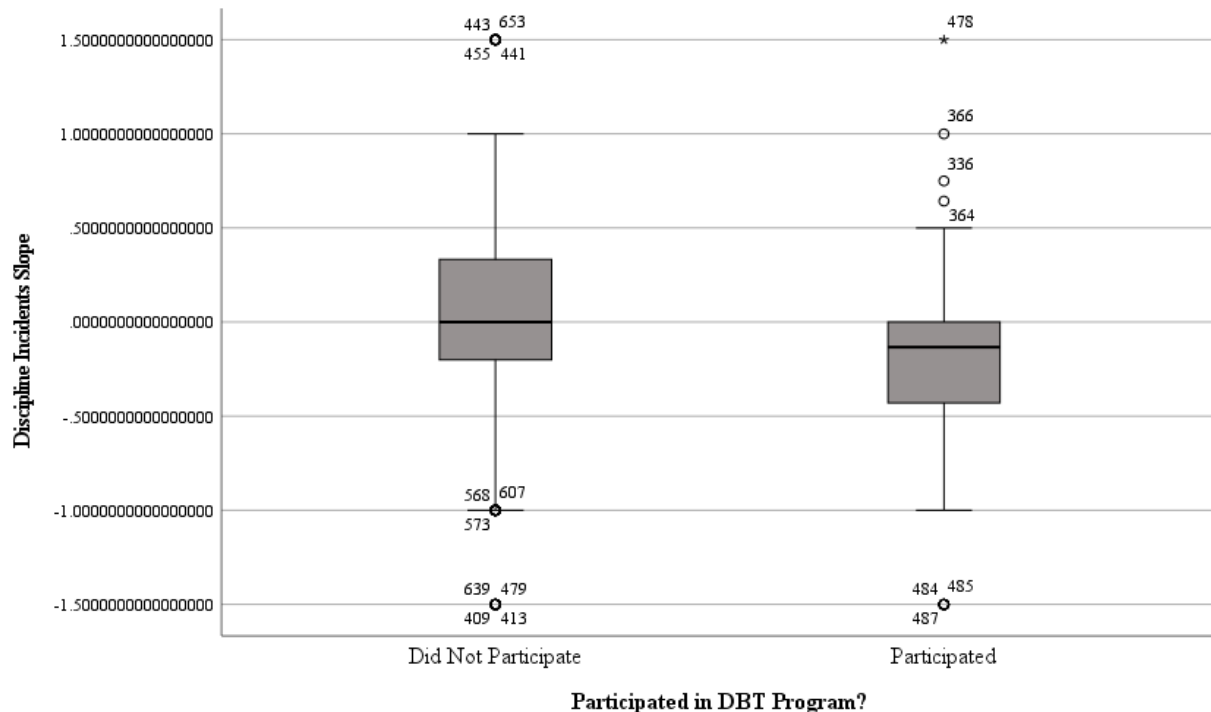
This figure thus shows a rather small range for the grey box and the whiskers—especially for those who did not participate. That group also shows a lot of outliers. Together, this suggests that the data for student self-reported GEC slopes among those who did not participate in the program has a high peak and long tails. The concomitant histogram also shows that:



Looking again at the box-and-whisker plots, we can see how much variability there is in these data. There are many data that could be considered as outliers and—seen in relation to the whole range of scores—the differences in the medians are not great. This means that whether an adolescent participated in the DBT program provides relatively little information about the self-reported changes in their executive functions—and that there is a lot more yet to be understood about it, given the great variability. *Something* is differentiating the students, but little of that is the DBT.

And yet, as we'll see below, the effect of the DBT is still significant. Which should put into perspective a bit what little it can mean for something to be significant.

Normally, I wouldn't include box-and-whisker plots in a Results section since they are so raw, but let's please add to our Word file that one for student self-reported GEC slopes by DBT participation as well as the box-and-whisker plot for discipline incidents over time by DBT participation:



Please also give them captions when you copy them into the Word file.

### 3.4.3 Copying a Figure to Word

1. Right-click on the figure, and choose Copy as
  - Choosing Image will paste it as is, including the changes we made to the chart's colors, etc., however, this cannot be further edited in Word
  - Choosing either Microsoft Office Graphics Object or EMF will paste it without that formatting, but in a version that can be edited in Word
2. Paste this into Word after the table of descriptives

### 3.4.4 Correlations

I find it often useful to first look at the bivariate (zero-order) relationships between my variables—both predictors and outcomes.

1. In SPSS, click on Analyze > Correlate > Bivariate
2. Select what you want, but please include Spec\_Ed, DBT, Economic\_Distress, All\_EFs\_SR\_Slope, and Disc\_Inc\_Slope which have some interesting associations
3. Under Options, make sure Missing values are Excluded pairwise. Here as much as anywhere, excluding listwise both removes useful information and likely biases the data we are looking at unless the data really truly are indeed [missing completely at random](#).
4. In the main dialogue box, make sure Flag significant correlations is selected. You can Show only the lower triangle; personally, I think this can be good to present to others, but I like to see both halves for myself.

5. In that main dialogue, we *could* select to also present Kendall's tau-b and/or Spearman rho.

If you only included those variables, the first few rows of the correlation matrix generated should look like this:

*Correlations*

		Special Education Status	Participated in DBT Program?	Economic Distress?	All Executive Functions Slope - Student Self-Report	Discipline Incidents Slope
Special Education Status	Pearson Correlation	1	-.044	.004	.026	.063
	Sig. (2-tailed)		.328	.931	.639	.307
	N	507	507	456	326	266
Participated in DBT Program?	Pearson Correlation	-.044	1	-.082*	-.165**	-.189**
	Sig. (2-tailed)	.328		.048	.003	.000
	N	507	670	578	326	357
Economic Distress?	Pearson Correlation	.004	-.082*	1	.168**	.036
	Sig. (2-tailed)	.931	.048		.002	.514
	N	456	578	578	325	329
All Executive Functions Slope - Student Self- Report	Pearson Correlation	.026	-.165**	.168**	1	.091
	Sig. (2-tailed)	.639	.003	.002		.262
	N	326	326	325	326	153
Discipline Incidents Slope	Pearson Correlation	.063	-.189**	.036	.091	1
	Sig. (2-tailed)	.307	.000	.514	.262	
	N	266	357	329	153	357

\*. Correlation is significant at the 0.05 level (2-tailed).

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

If you added in other variables, the correlation matrix may be huge; in any case, they often will be for whatever analyses you do—especially exploratory ones. Because of this, I find it easier to explore them in a spreadsheet, so:

1. Right-click on the correlation matrix that's generated in the output and Copy As an Excel Worksheet (BIFF)<sup>9</sup>
2. In Excel, paste this into an empty sheet
3. Also in Excel, single-left click on the first cell with a correlation value
4. Under the View tab, click on the Freeze Panes button near-ish the middle of the ribbon. This will "freeze" all cells above and to the left of that cell, letting you scroll around the matrix while keeping the labels for the rows and columns visible.
5. You'll see that the correlation between changes in student self-reported GEC ("All Executive Functions Slope - Student Self-Report") and DBT is smaller ( $r = -.165$ ) but significant ( $n = 326$ ,  $p = .003$ ). The correlations between DBT participation and changes in "Discipline Incidents Slope" is also significant ( $r = -.131$ ,  $n = 124$ ,  $p = .020$ ). The correlation between "Discipline Incidents Slope" and self-reported GEC isn't significant ( $r = .091$ ,  $n = 153$ ,  $p$

<sup>9</sup>I am advocating copying out this table differently than the table of descriptives, above. You can use either method for either, but each of these seems to work best and use the fewest steps for their particular type of table.

= .262), but its association with DBT may make it affect the DBT – student GEC slope relationship.


Please add this table as well to the Word file after the box-and-whisker plot, and give it a caption, too.

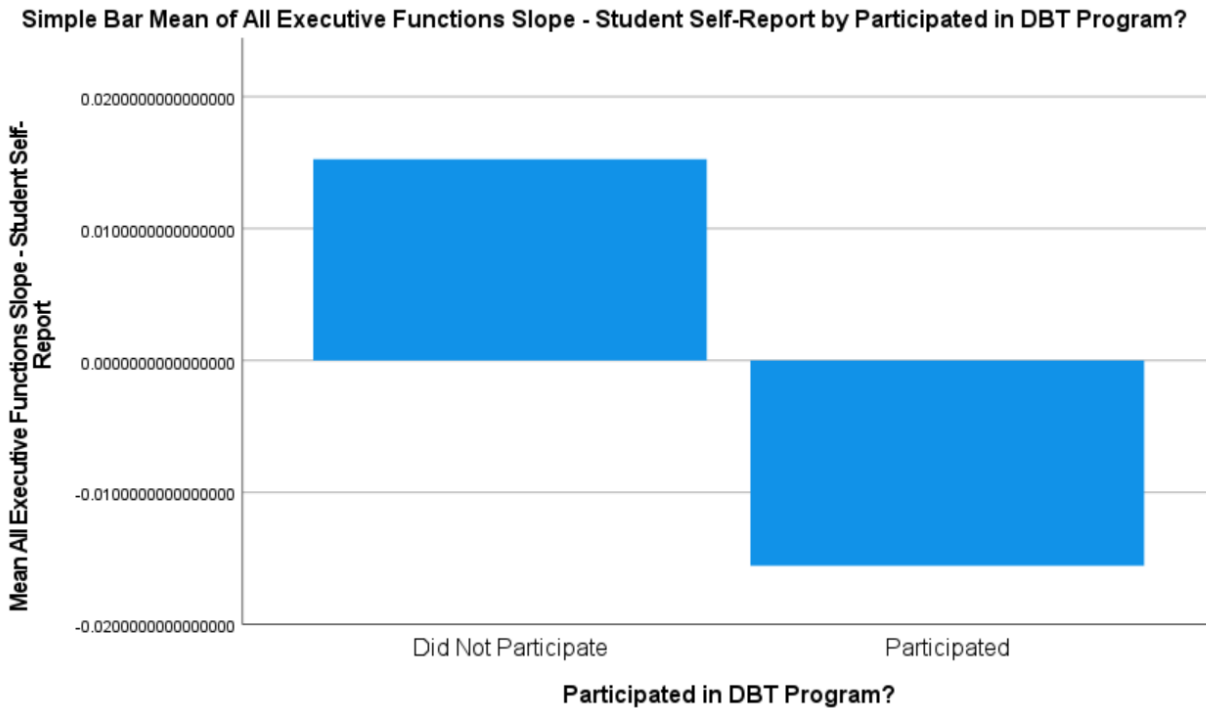
#### Note

Please remember that the instrument used to measure executive functions here, the **BRIEF**, is keyed so that *lower* scores denote *greater* executive functioning. (It asks respondents to report how often problems with executive functions happen, so higher scores denote more problems.) I nearly reversed-scored it here so that higher scores on all variables denote better things. However, I think it's a useful exercise to break the habit of relying on that. One reason I say that is in response to finding many beginning writers of research rely on using “positive” to indicate “good” things (“the intervention had a positive effect on recovery”) and “negative” to indicate “bad” things. Although this works fine for ordinary life, it is not defensibly appropriate for research since (1) it implies an inappropriate value judgement being made by the researcher and (2) there certainly are cases where positive numbers indicate less—or less good stuff. Higher values of blood pressure, BMI, A1C, antinuclear antibodies, etc. are often “bad,” and it would be confusing to say that something had a “positive” effect on their levels and then show negative correlations.

## 3.5 Creating Figures in SPSS

Explore creates several useful figures on its own, but SPSS has a pretty good general functionality to create a range of figures. Let's review a simple but common example.

1. Select Graphs > Chart Bulder
2. A dialogue will open asking you to ensure that all variables are put on the correct response scales; you can have this not appear, but I like to leave it as a warning to myself. Here, you can simply click OK
3. From the Gallery, choose Bar and select the first, simplest option: 
4. Drag that simple bar-graph icon into the chart preview window
5. Now, drag All\_EFs\_SR\_Slope to the *y*-axis and DBT to the *x*-axis to generate this figure:

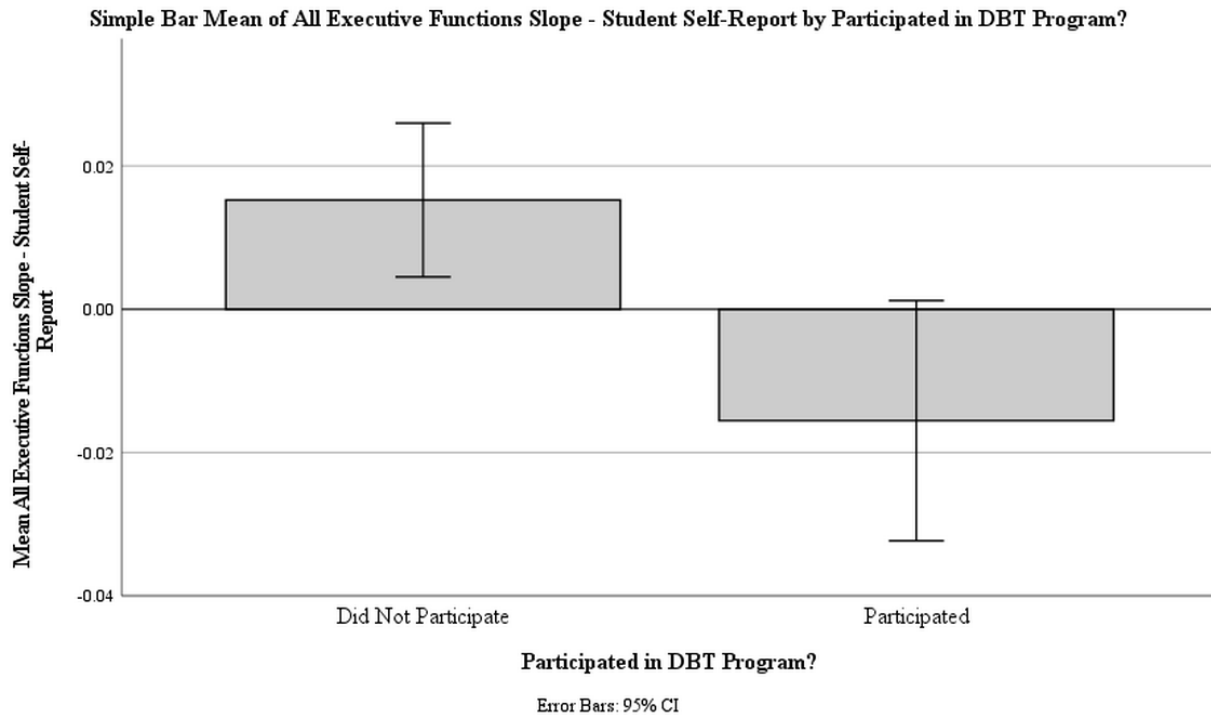


We can see that see the differences between the DBT groups much more clearly—but at the expense of a lot of information we know is there about the variability of the slopes.

But first, let's tweak the figure some more:

1. Under the Elements Properties tab to the right, select Bar1; we can make changes here to all of the bars. Here, we're simply going to select to Display error bars that are Confidence intervals of 95%
2. Click OK
3. Double-click on the chart
4. Single-right click on the numbers in the *y*-axis
5. Choose Properties from the window that opens (or type Cntl/Cmd + T)
6. The default range ("scale") of the chart is all right, but it's easy to change that under the Scale tab
  - Although I would prefer to select Display line at origin here since the origin here is simply zero—that there was no change in student GEC scores
  - Times when *would* change the range for the *y*-axis are when I have more than one chart
7. Select the Numbers tab, and change the Decimal Places to 2; click Apply—even before moving to another tab in this dialogue.

The chart should now look something like this:



Note that the 95% confidence intervals suggest that the self-reported executive functions of students who did not participate in the DBT program became significantly worse over these years (i.e., significantly more positive) since the 95% confidence interval for that group did not overlap zero. The change in executive functions for those who did participate overlaps zero, so we can't say, with sufficient confidence, that there was a change in those (at least not without taking into account other factors).

## 3.6 Analyzing Predictors of Self-Reported GEC Slope

### 3.6.1 First Model

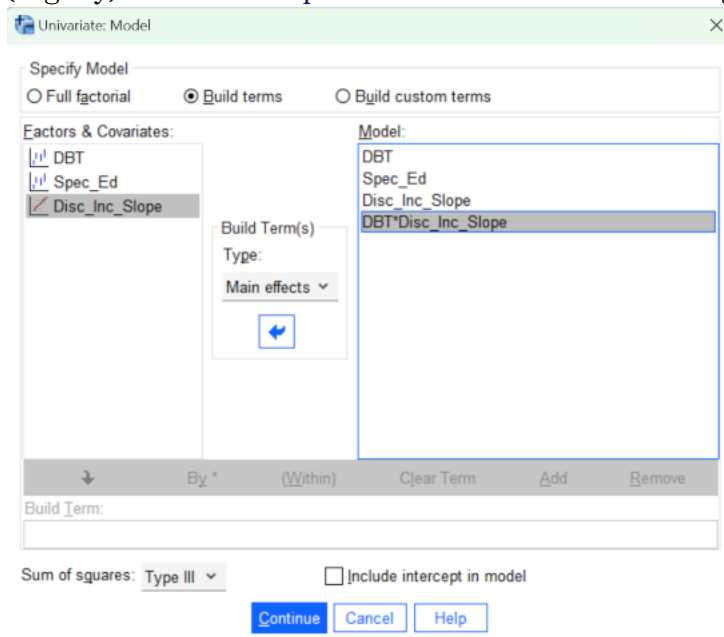
1. In SPSS, select Analyze > General Linear Models > Univariate<sup>10</sup>
2. Add All\_EFs\_SR\_Slope to the Dependent Variable field. We are seeing how well we can predict a teen's growth of executive functioning, so this is our outcome.
3. Most critically here, we're interested in whether participating in the DBT program changed—improved—the development of these teen's executive functioning. So, add DBT to the Fixed Factor(s) field. **Usually** any nominal variable can be placed in there.
4. Add Spec\_Ed as well to the Fixed factor(s) field.
5. Place Disc\_Inc\_Slop in the Covariate(s) field<sup>11</sup>
6. Click on the Model tab. When using General Linear Models (GLMs), you will nearly always go here to make sure it looks all right and to modify your model. The default is to analyze

<sup>10</sup>Statistics terminology is already unintuitive and complicated, and SPSS doesn't help by using terms that are sometimes uncommon—not wrong, but not ones you'd expect. “General linear model is the correct and common term, but by “univariate,” they mean only one outcome (“dependent”) variable; “multivariate” tests more than one outcome at a time, helping somewhat to control for correlations between the outcomes.

<sup>11</sup>Here is an other place where SPSS *tries* to be helpful, but—to me—fails. Random Factor(s) *ought* to be for continuous variables, but given how SPSS handles them, they're better placed under Covariate(s) regardless of how you interpret them.

the Full factorial model; this includes all main effects and interactions—even higher-order interactions (like 3- and 4-way interactions) that are usually futilely hard to interpret let alone communicate. Instead, select to Build terms.

7. Under Build Term(s) (in the middle, between the Factors and Covariates and Model fields), select Main effects
8. Now, under Factors and Covariates select all of the terms (single-left click any, then Cntl/Cmd + A) and move them via that arrow to the Model field.
9. Now change Build Term(s) to Interaction. Disc\_Inc\_Slope correlates with both All\_EFs\_SR\_Slope and DBT, so it *may* moderate the relationship between those two; given this, let's add the Disc\_Inc\_Slope *times* DBT interaction
10. The terms are all either dummy-coded or standardizes (into z scores), so we can de-select Include intercept in model and pat ourselves on the back for using our stats-fu to (slightly) increase the **power** of our model. This dialogue box should now look like this:



11. The Contrasts of our model are all right since the dummy-coding will do all of that for us.
12. Although we're not adding any Plots, they can be useful to investigate relationships among predictors.
13. The EM Means are estimated marginal means. These can be useful for investigating our well our model fits our data, so let's select to compute them for (OVERALL) and DBT. Residual Plots are also informative, so please select that as well.
14. Under Options, please select estimates of effect size, an option that really should be selected by default.
  - Descriptive statistics is just SPSS trying to make you think it's doing a lot when it's just doing the same thing twice.
  - Observed power is nearly always useless, even if **some** think it means anything.

### 3.6.1.1 Source Table of the First Model

The Test of Between-Subjects Effects table summarizes the results of this initial model:}

### Tests of Between-Subjects Effects

Dependent Variable: All Executive Functions Slope - Student Self-Report

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Model	.056 <sup>a</sup>	5	.011	1.987	.084	.063
DBT	.005	1	.005	.971	.326	.007
Spec_Ed	.003	1	.003	.456	.501	.003
Disc_Inc_Slope	.009	1	.009	1.650	.201	.011
DBT * Disc_Inc_Slope	.008	1	.008	1.339	.249	.009
Error	.836	148	.006			
Total	.892	153				

a. R Squared = .063 (Adjusted R Squared = .031)

The parts of this table are:

- Source
  - The sources of variance—hence why this is called a source table. Here, these include:
    - \* **Model:** The total amount of variance in the data that is accounted for by the whole model, i.e., all of the variables and any interaction terms
    - \* **DBT** etc: These are the individual terms in the model, one row for each term. Note that the DBT\*Disc\_Inc\_Slope row presents the results for the DBT × Discipline Incidents Slope interaction.
    - \* **Error:** This is the variance in the data not accounted for by the model.
    - \* **Total:** The total amount of variance in these data. Data sets with more variance have more information—more to learn, but more to figure out and account for.
- Type III Sums of Squares
  - We are computing a linear regression (using a general—i.e., not specified/specialized—formula), which—of course—means that a line is plotted through our data<sup>12</sup>. Unless the model (and the line drawn by it) perfectly fit our data, then each individual's data point may be some distance off this line. If, overall, the data points fall far from the line, then that means that the line (and the model we're testing based on it) don't account for much of whatever's going on in the data. We compute these distances from the line by squaring their distance; we then add up—sum—these squared distances. This is the Sum of Squares.
  - There are, in fact, [different ways](#) to compute this sum of squares, but the third way is currently the most common (and has been for a long time); it is more often used because it can more flexibly account for a wider range of models with different numbers of main and interaction effects.
- df
  - These are the degrees of freedom need to estimate the effects of each term (or, in the Error row, that are left over after estimating the model; and, in the Total row, the total number available in the model).

<sup>12</sup>The *y*-axis of this plot is the outcome variable (here, changes in student-reported executive functioning). The *x*-axis is, well, there are one axis per term in the model (here, 4 terms), so it gets pretty hard to visualize.

The idea is this:

Again, there is only so much information in any set of data. One way of thinking about that information is that every person brings a piece of information. We want to convert that “raw” information into insights. However, there are only so many insights we can make, of course. Each insight costs us. How much does an insight cost? Each insight—each numerical value computed to describe an effect—costs one degree of freedom<sup>13</sup>.

How many “insights” are needed to determine the effect of a given variable/term in a model? Well, it depends on how complex a variable/term is. Determining what’s going on in a variable that itself has a lot of levels costs more than it does to see what’s going on in a simple variable with only a few levels. If, for example, we want to understand what the effect of race is on something, then we actually want to figure out what the effect of *each* racial category is, and so each level will cost us. Each value for each race would need to be calculated ... almost. In fact, we get the estimate for one level for “free” because one level’s value is the default level estimated by the overall model; we really just need to determine how much each of those other levels differs from the default level. This “default” level is called the reference level since all other level’s value are given in reference to that one level. Which level we (or the model) sets as the reference only matters if there’s a theoretical reason<sup>14</sup>; otherwise it doesn’t, it just saves us a degree of freedom.

A variable/term that only has a few levels takes less to estimate. A dummy variable, with only two levels (has/hasn’t; is/isn’t; etc.) in fact only needs to make one insight: The overall model already estimates what it’s like *not* to have/be whatever the dummy variable is measuring. I mean, that’s technically true for *everything* not included in the model, isn’t it? The effect of, say, having freckles isn’t estimated in the model—the issue is when we want to estimate the effect of *having* freckles. Of course, we can only estimate the effects of factors that were measured, but it is nice to know that measuring effects that are dichotomized as dummy variables makes for a very efficient way of finding insights<sup>15</sup>.

- Mean Square

- This is how much information there is in a variable/term divided by how many “insights” that information has to figure out. If a variable has to estimate several values but doesn’t have much information to use to do so, then there likely isn’t much there to worry about<sup>16</sup>

Here, then, the whole model’s Sum of Squares is .056, but that information is divided up among four terms, and we must also spend a degree of freedom to estimate the default level for each variable<sup>17</sup>; we thus must divide that .056 up five ways, and  $\frac{.056}{5} = .011$ , the Mean Square value for the whole model.

<sup>13</sup>They’re called “degrees of freedom” because they do represent the number of insights we are free to make. After each insight is made—after each values is estimated—we’re left a little less free to make additional insights.

<sup>14</sup>To continue with the race example, it could be that we’re interested in seeing if, say, Blacks are treated differently than members of other racial categories; in this case, the value for the treatment of Blacks would be the reference level and the values for all others would be given in relation to how Blacks are treated. The values for the others, therefore, would really be for being treated worse/better than Blacks.

<sup>15</sup>“But if we divided up, say, race into a series of dummy variables, doesn’t that still take as many degrees of freedom to estimate all of those dummy variables as it does to estimate the effects of one ‘race’ variable with that many levels?” Yes, smarty pants, you’re nearly right. The thing is that we must first use all of those degrees of freedom to estimate the effect of that one, multi-level ‘race’ variable, and *then* run a post hoc test to see where any differences are, making us run an additional hypothesis test that (re)opens us up to making a Type 1 error.

<sup>16</sup>It’s the *mean* square because it’s the mean of the sum of squares—the average amount of information (deviance from the regression line) in each “insight” made by each degree of freedom.

<sup>17</sup>Luckily, we only need to spend on degree of freedom to estimate the default level for all of our variables/terms, no matter how many we have.

- F

- This is the ratio of how much information there is explained per degree of freedom in a given term relative to how much information is there left to be explained in these data. It's the Mean Square for a given term divided by the Mean Square for the Error row. I.e.,

$$F = \frac{\text{Mean Square}_{\text{Term}}}{\text{Mean Square}_{\text{Error}}}$$

Looking at the Model term again, the Mean Square is .011. The Mean Square for the Error term is .006. So the Mean Square for the model is about double the Mean Square for the error term. Specifically,  $F = \frac{.011}{.006} = 1.987$ . This means that, on average, the “insights”—the piece of information in each degree of freedom—made by our model account for about twice as much information as any other “insight” made by any other piece of information that could have been gleaned from our model. Is that good? Well .

..

- Sig.

- Presents the significance level of that term. This is usually (i.e., outside of SPSS) referred to as  $p$ , the probability of a false positive (Type 1 error). Convention in the health & social sciences, of course, is to tolerate making a false positive error 5% of the time. I.e., the  $p$ -value—the Sig. column—should have a value  $\leq .05$ .

So, no, explaining about half of the information in the data isn't enough. The Sig. value is .084—pretty close, but not significant. Since the overall model isn't significant, none of the individual terms should be either. In the *very* off chance any were, we wouldn't be justified to use them since the overall model itself isn't.

- Partial Eta Squared

- An ascendant idea in statistics (and thus quantitative research in general) is that of effect size. Significance (Sig.) determines *if* an effect is significant, but not *how* significant. Heck, one could argue that “significance” itself doesn't really matter—and there are certainly many cases when it doesn't. Instead, there is a growing conviction to instead make decisions based not only on significance of an effect but also (or instead) on the **size** of it.

Partial Eta Squared (or **partial  $\eta^2$** , the lower-case Greek letter “eta”) is a measure of effect size. It is “partial” in the sense of a partial correlation: It's the effect, e.g., participating in the DBT program after isolating that effect from the other terms in the model (i.e., of partialing out the other effects)<sup>18</sup>. (Cf. Chapter 6: Effect Size.)

There are different ways to compute effect size, but perhaps the simplest is also what's (essentially) used here. The effect size measure of, e.g., the DBT program is simply the difference between the EF slopes of DBT participants and the EF slopes of the non-participants. Like usual, though, we then standardize this mean difference so that we can compare effect sizes across outcomes and even across studies. And like is often done, we standardize it by dividing it by the standard deviation. So, the formula here is actually:

$$\eta^2_{\text{DBT Program Participation}} = \frac{\text{Mean}_{\text{EF Slope of DBT Participants}} - \text{Mean}_{\text{EF Slope of DBT Non-Participants}}}{SD_{\text{DBT Variable}}}$$

Cohen (1988) laid the foundation for defining effect size and on establishing *general* criteria for gauging how big is big (and how small is small). He suggested that for  $\eta^2$  (partial or not):

<sup>18</sup>So, what's partialled out when computing the  $\eta^2$  for the whole model? Nothing. The value in *that* cell is not a *partial*  $\eta^2$ , but in fact the whole  $\eta^2$  for the model—the size of the effect of everything therein. SPSS just “cheats” and puts the  $\eta^2$  for the model in the same column as the *partial*  $\eta^2$ s for the various terms.

- \*  $\eta^2 \leq .01$  is “small”
- \*  $\eta^2 = .06$  is “medium”
- \*  $\eta^2 \geq .14$  is “large”

It should—needs to be—noted that Cohen meant these small, medium, and large monikers to be suggestions—*not* absolutes. This is important to note since Cohen’s suggestions are already being ossified as rules that must be used. Large is large if that’s big enough to matter. Small is small if, well, O.K., 1% total variance is small, but still can matter especially for persistent effects.

A few more things to note about this source table. First is that the R Squared noted at the bottom.  $R^2$  is *also* a measure of effect size, and one you’ll see reported often. It’s the effect of an entire model, and also expressed as a proportion of total variance. Therefore, this  $R^2$  of .063 means that this model accounts for 6.3% of the total variance in the model<sup>19</sup>. Cohen would suggest that this is a “medium” size effect—even if it is not a significant one.

Second, note that the  $R^2$  for the model is the same as the “Partial Eta Squared” for the whole model. This will always be the case (within rounding), and both say that 6.3% of the total variance is accounted for in the model<sup>20</sup>.

Third, SPSS also provides an Adjusted R Squared parenthetically after the un-adjusted one. This is the  $R^2$  for the model after adjusting for the number of terms in it. We can improve the fit of a model simply by adding nearly any other variables from the data to it. Even if the newly-added variable doesn’t account for a significant amount of the variance, it will still account for *some*. Add enough non-significant terms and eventually the whole model itself will reach significance, even though no term within it is significant (or has a good effect size). To compensate for this, SPSS offers this adjusted value, which is reduced a bit for each term in the model, regardless of whether that term was significant or not.

Fourth, purposely, there are no significant effects. All of the  $p$ -values (Sig.s) are greater than .05. Let’s play further with it to see what we can find.

### 3.6.2 Second (Final) Model

Let’s remove the Disc\_Inc\_Slope  $\times$  DBT interaction. Interaction terms are often harder than main effects to find as significant, if for no other reason that the main effects—with which they surely correlate—are often added to the model as well, so the interaction term and the respective main effects are “fighting over” some of the same variance. This turns out not to do anything, but when we then also remove the Disc\_Inc\_Slope and SelfRefl\_Slope main effects, we find that the DBT main effect becomes significant. (Also remove Disc\_Inc\_Slope and SelfRefl\_Slope from the Covariate(s) field or SPSS will yell at you.)

The new model’s source table should look like this:

<sup>19</sup>What?! What about that  $F$  of  $\sim 2$  meaning the  $d$ f’s in the model give about twice as much insight as any other estimate we could make. The  $F$  score talks about the value of each insight (parameter estimate) and how good those are relative, essentially, to any other random estimate we could do on the data (relative to, say, the mean of odd rows is different from the mean of even rows). The effect size measures look at changes in the outcome variable, and whether the different levels of a parameter have different levels of the outcome—whether DBT participants had different EF slopes than non-participants. So, they use the source of information (the variance in the data), but are using it to answer different questions.

<sup>20</sup>And, again, although it’s in the Partial Eta Squared column, the value for the whole model is not a partial  $\eta^2$ , but the “full”  $\eta^2$ . SPSS just put it in the same column as the partial  $\eta^2$ s to conserve space and confuse newbies.

### Tests of Between-Subjects Effects

Dependent Variable: All Executive Functions Slope - Student Self-Report

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Model	.078 <sup>a</sup>	3	.026	3.796	.011	.034
DBT	.062	1	.062	8.998	.003	.027
Spec_Ed	.001	1	.001	.200	.655	.001
Error	2.213	323	.007			
Total	2.291	326				

a. R Squared = .034 (Adjusted R Squared = .025)

Removing those variables made for a significant model: The  $p$ -value is .011, which is less than .05. The effect size for the model is smaller, but we have narrowed ourselves down to a smaller set of variables that do explain significantly more about these teens' neurocognitive growth than most other things do. Looking at the terms in it, we see that it's DBT program participation that's doing the heavy lifting.

Were we simply exploring these data to find a parsimonious model that accounts for the most variance, we'd probably remove the Spec\_Ed term as well. This term indicates whether a teen has been diagnosed with a special need (whether they're in special education), and many of the disabilities that lead to this designation are cognitive, so it seems informative and theoretically relevant to leave it in and show it's weak effect here.

Why did the model (and the DBT term in it) become significant? Looking back at the correlation matrix, we see that Discipline Incidents Slope and DBT participation were correlated ( $r_{pb} = -.189$ ,  $p < .05$ ). Since DBT and Disc\_Inc\_Slope shared a significant portion of their variances, partialing out Disc\_Inc\_Slope removed a significant portion of DBT in our model, leaving less of it to show a significant relationship with All\_EFs\_SR\_Slope. This is not unusual, especially in field-based studies where we can't create well-controlled (and simpler) situations.

Note, however, that participation in the DBT program was essentially random. It cannot be that Disc\_Inc\_Slope, which measure whether a teen is getting in a increasing/decreasing amount of trouble at school, affected whether they participate in the DBT program. It *could* be that participating in the DBT program affected the number of times they were disciplined for getting into trouble, though. When these data were collected, they had only participated in the DBT program for two years, but if the program affects the development of their executive functions, it may well also affect their propensity to act out. This would make for a great [structural equation model](#) testing the [causal relationships](#) between those three variables.

Let's do one more thing to this table, and then move on to writing parts of a Results section.

1. Right-click on the table and selection Edit Content
2. Choosing either to edit it In Viewer or In Separate Window, right-click again on that table.
3. Choose TableLooks...
4. In the dialogue that opens, select APA\_TimesRoma\_12pt from the list of TableLooks Files. Note that you can Reset all cell formats to the TableLook; this will make any further tables you create also have this formatting. The table should now look like this:

*Tests of Between-Subjects Effects*

Dependent Variable: All Executive Functions Slope - Student Self-Report

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Model	.078 <sup>a</sup>	3	.026	3.796	.011	.034
DBT	.062	1	.062	8.998	.003	.027
Spec_Ed	.001	1	.001	.200	.655	.001
Error	2.213	323	.007			
Total	2.291	326				

a. R Squared = .034 (Adjusted R Squared = .025)

1.

Right-click once again on the table and select Copy and then paste it into the Word file where it will now be a Word-style table.

## 3.7 Writing the Results

### 3.7.1 Overall Strategy

My general strategy for organizing Results sections is to:

1. Think of the main points I want to make in my Results. This usually revolves around the research questions (hypotheses, whatever) that are the goals of the manuscript;
2. Create a set of visual displays (tables and figures) that present the main points I want to make;
3. Orient the reader to the content of those visuals,
4. While focusing on the parts of those visuals that relate to my main points,
5. And supporting what I describe in those visuals with stats (that are usually given only parenthetically).

I can and do discuss other results, both that are presented in those visuals and otherwise. But, I try to maintain the focus of the conversation on the main points of the manuscript. So, if there are additional points that are interesting / need to be explained but aren't directly related to the main points, I will indeed discuss them. However, I will try to make it clear that those are indeed ancillary points. I will do this sometimes by saying so.

Often, however, I try to separate the main points from ancillary ones in how I organize the Results. This may simply mean relegating the ancillary points to separate paragraphs. This works especially well if you do indeed ensure that each paragraph has a clear topic sentence, and that each other sentence therein directly relates to—usually simply expands upon—that main sentence. It's also often best to place the topic sentence first.

Separating out ancillary points may, however, also mean creating subsections of my Results, perhaps into an **Ancillary Analyses** section, if a more explanatory heading doesn't fit. This is an other time when templates can help, since you need only format a subheading, give it a clear title, and keep on typing. I may be a bit over-zealous in my use of subheadings, but I've yet to

have anyone complain. to the contrary, I've heard that it helps both to understand what it being discussed and to help readers find the topics they're currently interested in. It sure helps me find them when I revise my writing.

### 3.7.1.1 MAGIC Arguments

In a complementary view, Abelson (1995) posits that persuasive, scientific arguments contain five, general, **MAGIC**al characteristics:

1. **Magnitude** pertains to Abelson's contention that "the strength of a statistical argument is enhanced in accord with the quantitative magnitude of support for its qualitative claim" (p. 12). Among the measures of magnitude are *effect size* and confidence intervals. I wholly agree these should be presented and considered in most Results sections. I find his presentation of "cause size" interesting but unnecessary.
2. **Articulation** refers to Abelson's recommendation to include as much detail and specificity about the differences (or lack) between groups. His formal presentation of articulation in [Chapter 6](#) is interesting, but I find this example he gives to be sufficient:

[an] investigator is comparing the mean outcomes of five groups: A, B, C, D, E. The conclusion "there exist some systematic differences among these means" has a very minimum of articulation. A statement such as, "means C, D, and E are each systematically higher than means A and B, although they are not reliably different from each other" contains more articulation. Still more would attach to a quantitative or near-quantitative specification of a pattern among the means, for example, "in moving from Group A to B to C to D to E, there is a steady increase in the respective means." The criterion of articulation is more formally treated in chapter 6, where we introduce the concepts of ticks and buts, units of articulation of detail (p. 11).

3. **Generality** is similar to generalizability in that it "denotes the breadth of applicability of the conclusions" (p. 12) but also contains elements of "triangulating" insights through multiple studies and methods. He argues that "[h]igh-quality evidence, embodying sizable, well-articulated, and general effects, is necessary for a statistical argument to have maximal persuasive impact, but it is not sufficient. Also vital are the attributes of the research story embodying the argument" (p. 12). He further expands upon his ideas in relation to an ANOVA in [Chapter 7](#).
4. **Interestingness** and credibility (covered next) are elements Abelson sees as important components of an effective research narrative. His coverage of it in [Chapter 8](#) is indeed useful. A quick summary of his ideas are:

that for a statistical story to be *theoretically* interesting, it must have the potential, through empirical analysis, to change what people believe about an important issue. This conceptual interpretation of statistical interestingness has several features requiring further explanation, which we undertake in chapter 8. For now, the key ideas are *change of belief*—which typically entails surprising results—and the *importance* of the issue, which is a function of the number of theoretical and applied propositions needing modification in light of the new results (p. 11).

5. **Credibility** "refers to the believability of a research claim. It requires both *methodological* soundness, and *theoretical* coherence" (p. 13). Although a bit antiquated, Abelson's discussion of this is expanded upon quite well in [Chapters 5](#) (of statistical errors) and [9](#) (of methodological ones). Perhaps it can suffice here to note that he says that:

[t]he requisite skills for producing credible statistical narratives are not unlike those of a good detective (Tukey, 1969). The investigator must solve an interesting case, similar to the “whodunit” of a traditional murder mystery, except that it is a “howcummit”—how come the data fall in a particular pattern. She must be able to rule out alternatives, and be prepared to match wits with supercilious competitive colleagues who stubbornly cling to presumably false alternative accounts, based on somewhat different clues (p. 11)

### 3.7.2 Writing Style

Some general advise<sup>21</sup>:

1. Reverse-engineer what you read. If it feels like good writing, what makes it good? If it's awful, why?
2. Let verbs be verbs. “Appear,” not “make an appearance.”
3. Beware of the Curse of Knowledge: when you know something, it's hard to imagine what it's like not to know it. Minimize acronyms & technical terms. Use “for example” liberally. Show a draft around, & prepare to learn that what's obvious to you may not be obvious to anyone else.
4. Avoid clichés like the plague (thanks, William Safire).
5. Old information at the beginning of the sentence, new information at the end.
6. Save the heaviest for last: a complex phrase should go at the end of the sentence.
7. Prose must cohere: readers must know how each sentence is related to the preceding one. If it's not obvious, use “that is, for example, in general, on the other hand, nevertheless, as a result, because, nonetheless,” or “despite.”
8. Revise several times with the single goal of improving the prose.
9. Read it aloud.
10. Find the best word, which is not always the fanciest word. Consult a dictionary with usage notes, and a thesaurus.

#### 3.7.2.1 Strong & Clear Organization

This [Stack Exchange question](#) addresses writing the Introduction.

#### 3.7.2.2 Strategies for Discussing Various Analyses and Research Arguments

Abelson's (1995) discussion in [Chapter 4](#) of rhetorical styles and how they relate to different types of data/analyses are readable and sometimes useful. Among the topics of especial interest there are his coverage of:

- [One- and two-tailed tests](#),
- [The roles of parametric and non-parametric tests](#),
- [Absolute versus relative effects](#),
- [Different ways to frame analyses](#), and
- [How to write about  \$p\$ -values](#).

---

<sup>21</sup>Unfortunately, I don't remember where I got this, but it's not mine.

### 3.7.2.3 Simple, Direct Sentences

Write simply, striving for language that **nearly anyone** could understand. Make your point fast, clear, and easily found.

Simplicity suggests using short, direct sentences. The traditional subject-verb-object (SVO) format is thus also often best; and yes, this *does* mean writing in the first person for yourself and in terms of what other researchers write and did. **APA 7th finally advocates doing so**, and it's about time.

Writing in subject-verb-object format has several advantages. First, it helps separate out ideas (and actions) that are yours versus those that are others. It thus lets you both give credit where it is due and distinguish subtle differences between perspectives. It makes it easier to show what is your own perspective separate from others, both to show what you are adding to the conversation and to help inadvertently give the impression you're taking credit where it isn't due.

Second, I do believe it will help clarify the sources of your ideas and evidence, thus allowing for—encouraging—a more sophisticated consideration of the various sources of those ideas and evidence.

Third, writing in SVO makes it a *lot* easier to write lit reviews *and* to avoid the novice phrasing of saying “The literature proves that life is good.”<sup>22</sup>

Writing is also mindful practice. Attend to what you are saying—not just to what you are writing. Attend to what it is like to read your writing<sup>23</sup>. You can practice this by doing the same with what you read. Notice when you do or don't like someone else's writing—and figure out how you can do that in your writing.

Yes, writing is re-writing. It is exceedingly rare that I've made a manuscript worse by reviewing it. In my opinion, writing and editing are not completely the same set of skills, and so we must practicing both sets to achieve our best writing. Critiquing others' writing can help develop our own editing skills, as can having others critique our own work and us then actively reflecting on their suggestions. However it works best for you to do it, I suggest you—we—work to develop the skill of revising as well as writing.

### 3.7.2.4 Mind the Reader

I indeed try to write as if I'm having a conversation. As should be true of any conversation, I try to think about who the other person is, what they know and don't, what parts of my topic are hardest to understand, and certainly what they likely want to know about the study.

It's safe to assume that in general most people aren't ever going to be as interested in what you write as you are. Don't assume they'll stay engaged.

The preponderance of your readers are going to be reader your writing to get a point. So make your point as clearly and simply as you can; help the readers find it fast.

And help them understand it. Please don't assume your reader understands your works just because you do.

---

<sup>22</sup>*Ugh!* Sure, we can consider “literature” to be a **metonym** standing in for the authors, but it's often easier to track the progress of science through those conducting it than through a series of disconnected articles. In addition, “the literature” is such an imprecise and lazy phrasing that nearly prevents any real consideration of the nuances between perspectives. And **please**, when was the last time *anything* was *proven* in science?!

<sup>23</sup>And yes, that means you have to read it. Which means rewriting it.

### 3.7.2.5 Verb Tense

The correct tense for discussing *writing* and *ideas* is the *present* tense; thoughts and words are considered to be “alive.” It also sounds better to say “Jones (2020) says life is good.” It also means you needn’t worry about *when* they wrote it.

*Actions*, however, are correctly relegated to the *past* tense. “Jones (2020) conducted a study of neonatal nurses and from that concludes that she believes life is good.” This does mean that you should put the actions *you* conducted in your study in the past as well (since you’re presumably writing after you did them; kudos to your multitasking mastery if you’re writing while you’re doing them!).

### 3.7.2.6 Follow Conventions

Some of the suggestions I’m presenting here are little more than repetitions (or re-packagings) of standard conventions. Some come from the surprisingly-useful and -readable Chapters 4, 5, and 6 of the 7th edition of the *APA manual*, the first two of which also address beyond simple writing style and grammar, such striving for bias-free writing.

Additional, useful guides to following conventions include:

- Desmarais, C. (June 11, 2017). 43 embarrassing grammar mistakes even smart people make. *Inc.* <https://www.inc.com/christina-desmarais/43-embarrassing-grammar-mistakes-even-smart-people-make.html>

## 3.8 Writing about These Results

Let us now begin to write about the results we’ve put into that Word document.

I generally write about results by first using descriptive statistics to orient the reader to the data. I also use this time to address any issues with them, how I either addressed those issues, and/or how those issues affect analyses and their interpretation.

I also think ahead of time about the main points I want to make in a Results section, and then both organize the information and orient the writing in the sections towards that. You’re essentially using the Results section to give evidence toward supporting a position you take about your data and what it says. Of course, this position ought to address your research goals/hypotheses.

This example was exploratory (exploring the data and seeing what we found that we thought was interesting—not testing an a priori research goal outside of “what’s up with the DBT”), but our exploration did note a few things of interest. To me, these include:

- There were a lot of missing data.
- There was a lot of variability within the data that we had.
- Changes in executive functioning (and in, e.g., discipline incidents) were often slight.
- The story of the relationships between the variables is relatively complex.
- Nonetheless, the zero-order<sup>24</sup> correlations suggest a few interesting relationships in relation to the DBT program.

---

<sup>24</sup>I.e., simple correlations between two variables—not part or partial correlations.

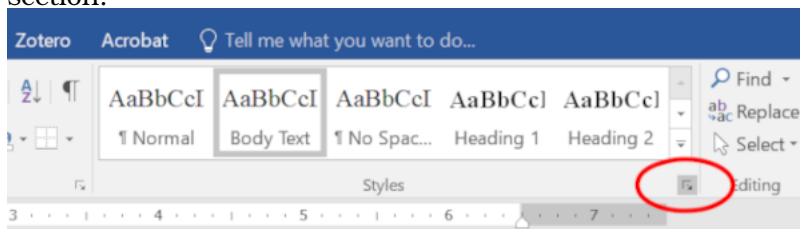
- A more structured investigation of those relationships—conducted also in light of a dash of theory—finds that participating in the DBT program was associated with changes in student-reported executive functioning.

Those are thus the points—and perhaps a reasonable order to address them in—to cover in this Results section.

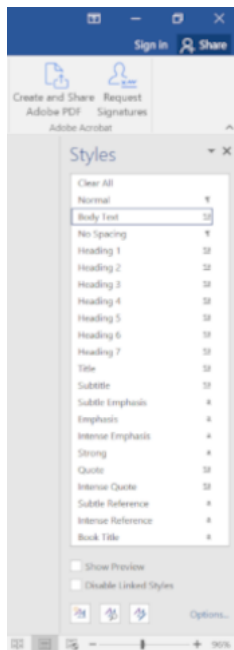
### 3.8.1 Descriptives

So, let's indeed start by creating a level 2 header entitled “Descriptives”:

1. In Word, go the Home ribbon, click on the tiny arrow in the bottom right corner of the Styles section:



2. A side menu should appear to the right:



3. Go the Results section, and—on its own line—write Descriptive Statistics
4. With your cursor anywhere in the line where you wrote Descriptive Statistics, click on Heading 2 in the Styles side menu. If you're still having issues, consult Chapter 14 where more was given about using a style template and reference manager, including how to instead use LibreOffice Writer<sup>25</sup> and RStudio.

<sup>25</sup>Where it is much easier to use styles.

With the Descriptive Statistics table just below where you're writing, start by mentioning that table to let your readers know outright that that's what you're talking about and where to look for more information.

E.g.:

Table 1 presents descriptive statistics for the main variables investigated.

Now, discuss some highlights noted in that table. Your goal is *not* to restate what is given in that table. For the most part, **either** present information in a table/figure **or** discuss it in the text.

Readers often expect to be quickly oriented to whom the data represent—who are these participants we'll be talking about throughout the Results, so that's a good thing to address near the beginning of the section, i.e., now:

As we can see in this table, 40% of these middle and high-school students were diagnosed with disabilities that warranted being in special education; 62% of these students were under economic distress.

Again, the outline we sketched for the Results section above included noting that there were a lot of missing data and that there was a lot of variability within the data we have, so in that first paragraph discussing this table we can write, e.g.:

However, it is important to note that although data were available on each variable for up to 670 students, we only had complete data on 152, reducing the generalizability of these findings—including to this particular population<sup>26</sup>.

On average<sup>27</sup>, there was little change in an overall composite measure of these students' self-reported executive functioning (mean change = .01<sup>28</sup>). The number of disciplinary incidents a student was part of tended to decrease slightly ( $\bar{x}$ <sup>29</sup> = -.04). However, for both of these variables, the standard deviations were much larger than the means, indicating that there was much variability in these students' development and behavior<sup>30</sup>.

And yes, I try both to assume that readers know less about the field than one would expect and to use this opportunity to restate or explain concepts. I also tend to write longer, more detailed descriptions both of tables and results early in the Results and then give briefer, more succinct descriptions later; experienced readers can easily skim over the more detailed descriptions given

<sup>26</sup>It is gauche to editorialize or give opinions in the Results section (those are for the Discussion). However, it can be quite helpful to explain what given results mean objectively both to help more novice readers and to make your point about what in the results you're focusing on.

<sup>27</sup>Starting new paragraph here makes for a few very short paragraphs, but it seems to me that the next topic of changes in EFs & disciplinary incidents is thick enough to warrant separate attention

<sup>28</sup>Somewhere earlier in the manuscript—probably in an Analyses or Analytic Strategy subsection of the Methods section, I would have given a careful and detailed explanation of what this variable actually measures. Still, it's worth giving some review here, both to remind readers and because, in my opinion, a well-written manuscript allows readers to read sections out of order—and even to skip some sections entirely. Nonetheless, it's a balance everyone must find on their own about how much to restate things at the expense of flow.

<sup>29</sup>It may be confusing to first say "mean change" as I did for EFs and now use  $\bar{x}$  for discipline incidents, so perhaps going with one style or the other is better.

<sup>30</sup>Although I do strongly advocate writing SVO-style sentences, it seems useful for this one to first explain what it is we're talking about before then discussing a slightly complex point.

earlier while more novice ones both gain more of a footing and then enjoy seeing themselves able to apply it later in the section.

Nonetheless, it can also be appropriate to write brief and direct pieces. Please add this sentence to the same paragraph:

About 20% of these adolescents participated in the DBT program<sup>31</sup>.

The percents etc. in that table are presented there separate from each other. In other words, the 357 students for whom we have discipline incidents are not necessarily the 326 students for whom we have executive functioning data. And normally the next table would indeed be to present how much overlap there is between those who have missing data in these variables<sup>32</sup>, but we're simply covering the overall strategy for writing Results sections, not all that should go in one.

Nonetheless, it is important to reminder readers of this—and that you are cognizant both of this yourself and of the implications this has on the generalizability of your results. Therefore, something like this is warranted:

It is important to note from Table 1 that we only have complete data for 152 (22.7%) of the 670 participants.

In fact, a general rule of thumb is that missing more than 10% of the data is problematic. The data we're using here has more missing values than the actually data since I didn't include it all, but even there this is an issue my colleagues and I must contend with. Make the above sentence about missingness a topic (and first) sentence of a new paragraph, and add a few more sentences explaining the ramifications of this degree of missingness on the results and their interpretation.

Another aspect about the executive function and discipline incidents variables is their relatively large variance. The table shows this through the magnitude of the standard deviations relative to the means, but I think it's easier to see this in the box-and-whisker plots we made through the data exploration. We could have talked about this earlier in our discussion of that table, but putting it last lets us segue to those figures. Or rather, since this insight is supported by both that table and those figures, it makes sense to talk about this aspect last:

Finally, the large standard deviations for the all executive functions and the discipline incidents slopes<sup>33</sup> relative to their respective means in Table 1 indicate that there is a lot of variability among these slopes. These adolescents displayed a wide range of developmental trajectories.

Again, I'm trying to point out important piece of information in the tables (and soon figures) and explain what it means *while also remaining objective*. Remember, for the Results section, just give the facts.

In that same paragraph about the variability, we can say:

<sup>31</sup>It may be clearer simply to say "19%" than "About 20%" since the former more clearly relates to the table; I went back and forth a bit on this.

<sup>32</sup>We could do this in SPSS by first re-coding the variables into dummy ones based on whether they're missing with (1) Transform > Recode into Different Variables, then (2) selecting System- or user-missing as the first Old Value and 0 as its New Value, and then (3) choosing All other values as the Old Value and 1 as its New Value. Once the variables are coded into missing/non-missing dummies, we can see how much overlap there is in the missing values by comparing these missing/non-missing dummy variables with Analyse > Descriptive Statistics > Crosstabs.

<sup>33</sup>Variables—really everything but the title—is to be put into title case in tables & figures. However, APA 7th (p. 169) has variables (and experimental conditions) put in lower case in the text.

This variability is perhaps better seen in Figures 1 and 2 which present box-and-whisker plots of the slopes in self-reported executive functions<sup>34</sup> for DBT program participants and non-participants (Figure 1) and slopes for discipline incidents (Figure 2).

After this, please describe what these figures say about the distribution of scores, including their variability relative to each other and the implications of this on the results & their interpretation.

I considered ways of organizing these results so that we could also use the box-and-whisker plots to transition to the bar chart showing the mean difference & confidence intervals<sup>35</sup>, but it works fine (I think) to instead discuss the correlations first and then use that as a backdrop to jump the mean differences and the linear models testing them.

In any case, I do feel it is important to discuss the correlations. First because it's often useful to present the bivariate relationships between variables to orient the reader to that next level of complexity to the data. Second because it helps us show where that complexity is, thus helping lay the foundation for interpreting models that include combinations of those variables. Indeed, as I mentioned in the section above conducting those analyses, the inter-relationships between the variables affects the significance of the terms in our models; both our readers and we are well-served to stay aware of these inter-relationships when interpreting the linear models.

Let's write an other paragraph about them:

The correlations between the variables presented are presented in Table 2. As we can see, the correlations between the variables all tended to be rather weak ( $r_s = -.004$  –  $-.189$ <sup>36</sup>). Despite these small values, participation in the DBT program correlated significantly with experiencing economic distress ( $r^{37} = -.082$ ,  $p = .048$ ), changes in discipline incidents ( $r = -.189$ ,  $p < .001$ ), self-reported changes in executive functions ( $r = -.165$ ,  $p = .003$ )<sup>38</sup>. Self-reported changes in executive functions also correlated significantly economic distress ( $r_{pb}^{39} = .168$ ,  $p = .002$ ).

It would help to explain to the their what each of these correlations mean—even if executive functions weren't scored counter-intuitively—so please do that after those sentences. For example, we could say:

<sup>34</sup>You may have noticed that I don't stick with a specific format for the variables—and *certainly* try to not divorce my writing more from the English my readers know by littering it with acronyms. Sometimes I feel it helps to keep the description of the variables in more flexible, lay terms, but sometimes it would help instead to use the same phrasing for them. I do go back and forth a bit trying to find what seems the best way of

<sup>35</sup>Something like, "These [box-and-whisker plot] figures suggest there may be some differences in the means of these these despite the great variability in scores."

<sup>36</sup>I considered giving the absolute values instead of showing that they're negative. This may have helped, and we easily could with "...rather weak ( $|r_s| = .004$  –  $.189$ )."

<sup>37</sup>Both DBT participation and economic distress are dichotomous (dummy) variables, so in fact a correlation between them is inappropriate. We should use a frequency table (like we get as a Crosstab) and instead compute either the Fisher's exact test or a  $\chi^2$  to test for significance. I'm punting on this here for expediency even though what I'm doing is wrong.

<sup>38</sup>I presented these three in a different order than they appeared in the table, I did this to make for a slightly easier transition to the next sentence that's also about self-reported EF slopes. However, were I to prepare this for publication, I would re-arrange the variables in that correlation matrix (and the table of descriptives) to make them all follow the same sequence as is given in the text.

<sup>39</sup>This is appropriate. We can correlated a dichotomous variable with a continuous one. (We can correlate continuous variables with anything else). A correlation between a dichotomous & continuous variable is called a "point biserial correlation" which is abbreviated as  $r_{pb}$ . The formula that produces a point biserial correlation is mathematically equivalent to the one that produces correlations between two continuous variables (a "Pearson correlation"), so including it in this table is perfectly legitimate. Still, I really should have, e.g., put "NA" for the cells presenting the DBT-economic distress correlation and instead included a frequency table for that.

Adolescents who were experiencing distress demonstrated significant decreases in their self-reported changes in executive functioning (although the correlation is positive, higher scores on the BRIEF indicate decreases in executive functioning)<sup>40</sup>.

We can then simply end that paragraph with:

All other correlations were not significant (smallest  $p = .262$ )<sup>41</sup>.

Next, please add a paragraph discussing the bar chart presenting the DBT group mean levels of changes in executive functioning. This is in fact a corollary of the point serial correlation between those two variables: The 95% confidence intervals for either group do not contain the mean of the other group (e.g., the confidence interval for the Did Not Participate group spans from about 0.03 to just a little over 0 while the mean for the Participated group is below 0.<sup>42</sup>). Part of the paragraph could read as either:

Those who participated in the DBT program self-reported significantly greater improvements in their executive functioning than those who did not participate ( $F_{3, 323} = 8.998$ ,  $p = .003$ ), however this effect was rather small ( $\eta^2 = .027$ ).

or:

Those who participated in the DBT program self-reported significant ( $F_{3, 323} = 8.998$ ,  $p = .003$ ) but rather small ( $\eta^2 = .027$ ) improvements in their executive functioning relative to those who did not participate.

### 3.8.2 Inferenceals

We can use our discussion of the difference presented in this bar chart to introduce the general linear model.

I asked you to only paste in the final model. Had this not been a rather atheoretical exploration of the data but instead a set of analyses comparing different theoretically-relevant configurations of the variables, then it may well have been appropriate to include them both. For example, the first models (the one we didn't paste in Word) included discipline incidents (and whether they interacted with DBT participation), so had we posited as an hypothesis that changes in discipline incidents would predict the effectiveness of the DBT program, then it would be worth showing that model and *then* also showing the model without it demonstrating that, although knowing changes discipline incidents didn't improve our understanding of changes in executive functions, information about them does affect our ability to understand the relationship between DBT participation and changes in executive functions.

But here, please simply take a crack at describing what is presented in the source table presented in Table 3. Please consider the results in light of the zero-order correlations between those variables and how partialing out having an IEP affects our interpretation of the relationship between DBT participating and changes in executive functioning. Please remember that there are many

<sup>40</sup>Courtesy Lisa Gillespie =^3

<sup>41</sup>Sure, we could also give the correlations along with the  $p$ -values, but this seems like enough to me, especially since there's nothing really there to talk about.

<sup>42</sup>As you may have guessed, yes, the fact that the Did Not Participate group's confidence interval does not overlap zero indicates that that group's slopes were significantly greater than zero: The members of this group experienced significant declines in their self-reported executive functioning over these years. Life is tough for these teens.

reasons these adolescents could have been given an IEP; not all disabilities are related to cognitive or emotional functioning (or information processing). In fact, even economics and race/ethnicity affect whether a student is diagnosed—and what sorts of diagnoses they receive.

### 3.9 Further Resources

- **General Writing Tips and Guides**

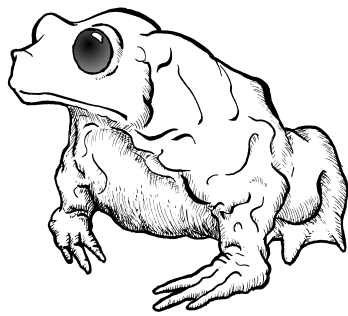
- Abelson, R. P. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.
  - \* The first couple of chapters are good; the rest starts to ramble.
- Cochrane, J. H. (2005). *Writing Tips for Ph. D. Students*
  - \* Readable and succinct, Cochrane puts a lot of good, general advice into a dozen pages.
- Katz, M. J. (2009). *From Research to Manuscript: A Guide to Scientific Writing* (2nd Ed.). Springer.
  - \* Several students have commented on how helpful this book is.
- *Statistical Writing*. UCLA: Statistical Consulting Group.
  - \* With some SPSS-generated output as talking points, this gives specific guidance for particular analyses.

- **Writing about Specific Types of Analyses**

- Checklists for results in general as well as various types of analyses:
  - \* Lang, T. A. & Altman, D. G. (2015). Basic statistical reporting for articles published in biomedical journals: The “Statistical Analyses and Methods in the Published Literature” or the SAMPL Guidelines. *International Journal of Nursing Studies*, 52(1), 5–9. <https://doi.org/10.1016/j.ijnurstu.2014.09.006>
- Mixed models and multilevel models:
  - \* *How should results from linear mixed models (LMMs) be reported?*, StackExchange Psychology & Neuroscience Community
  - \* Monsalves, M. J., Bangdiwala, A. S., Thabane, A., & Bangdiwala, S. I. (2020). LEVEL (Logical Explanations & Visualizations of Estimates in Linear mixed models): recommendations for reporting multilevel data and analyses. *BMC Medical Research Methodology*, 20(1), 3–3. doi: /10.1186/s12874-019-0876-8
- Meta-Analyses
  - \* The PRISMA model for reporting systematic reviews and meta-analyses

- **Guide for Simpler, Clearer Writing**

- The Hemingway App lets you enter/paste in text and then suggests ways to make that text better 8 **Very General Writing Tips & Guides**
- Desmarais, C. (2017) *43 Embarrassing Grammar Mistakes Even Smart People Make*. [Inc.](<https://www.inc.com/>) Life.





## Chapter 4

# Introduction to Measuring Relationships and Building Models

One of my goals in this curriculum is to give you a strong foundation in what I've learned to be among the principles and practices guiding major, currently-recommended analyses. As you may have also realized, doing this differs somewhat from how statistics is usually taught. I very sincerely hope my gamble is well made, and that you not only gain the same skills that others would gain from doctoral statistics courses, but that you magnify that with a good understanding of what the heck is going on.

To this end, I have tried to convey the importance of a few, inter-related concepts that permeate much of modern analyses.

### 4.1 Measuring Relationships

In a very readable [ScienceAlert article](#), Neild describes a study in which researchers found that, although night owls tend to have shorter lives than morning larks, being a night owl *per se* doesn't increase mortality in older adults. Instead, it was the riskier behavior these people of the night prefer more to do. But how could the researchers determine this? If pretty much all night owls didn't live as long, how could they say that being a night owl didn't increase mortality? What sort of analysis would allow them to remove—isolate—the effect of being a night owl (more technically, having a nocturnal chronotype)?

Isolating an effect begins with being able to measure the effect—more specifically, being able to measure the main effect of a variable (here, chronotype—early birds vs. night owls). Measuring it so allows us to circumscribe its effect—to locate and measure the size (and location) of its effect. That then lets us isolate the effect of it on other variables. We can “partial” out the effect, letting us see *other* relationships with that effect removed.

We must thus be able to measure the extent of these relationships well. The better it is measured, the better we can detect its boundaries and thus isolate it. Estimating the relationship between variables thus not only allows us to see how much they are related, but to isolate that relationship from *other* relationships.

How much two variables are associated with each other is often modeled as a linear relationship where the slope of the line (when the two variables plotted on the axes) shows how much they're

related: The greater that slope (the more it deviates from a flat line with a slope of zero), the stronger the relationship between them. We often use some assumptions to estimate the underlying nature of the variables, and then use some criterion (e.g., being 95% sure we're right) based on those assumptions to decide whether we think the association between those two variables matters—whether it is “statistically significant.”

If the two variables are measured on the same scale—say z-scores—then the greatest relationship would be represented by a straight line going up (or down if it's a negative relationship) at a  $45^\circ$  angle; whenever the line goes one unit to the right, it also goes exactly one unit up. The slope is  $\frac{1}{1}$ , or simply 1. Anything less than a perfect relationship results in a slope that is less than 1.

If we make no further assumptions about the two variables than that they have a linear relationship and that both sets of data are roughly homoscedastic, then we can compute a correlation coefficient. The exact method we use depends on the measurement level (dichotomous, ordinal, interval, etc.) of the variables, but all correlations are set to range between 0 and 1 (or -1) by convention.

So, one way to look at a correlation is the strength of the (linear) relationship between variables.

## 4.2 Signal-to-Noise Ratios

Another way to look at correlations is that they are one of the ways of measuring a “signal-to-noise” ratio—another idea that permeates much of statistics. Here, it's not just a question of how much two variables are associated, but how much that association accounts for all the information we have in our data about those variables—how much of the total variance is covariance between those things. The amount that the variables move together—the amount they covary—is a measure of the strength of their relationship. The amount they move independent of each other—the amount they do not share variance—is a measure of how weak their relationship is. If they have a weak relationship, then we obviously don't have a very good representation of what makes these two variables take on whatever values they have.

## 4.3 Building Models

The models we're analyzing in the class activities—that, say, participating in a DBT program is linearly related to growths in executive functions—are not very good; they don't account for much of the total variance in the data. Nonetheless, building models and testing how well they account for our observations (e.g., the data we have on hand and future observations we will make) is a third idea that permeates much of statistics.

Let's go back to our example of a very simple model: a correlation between two variables. Let us also say in this example that it's a weak correlation, and we want to improve on this model to make it a better representation of what's really going on in our data. We could do this by utterly throwing out that first model and trying another one (e.g., by seeing if another zero-order correlation is stronger). Or we could try tweaking our model, say, by adding in another, third variable (i.e., testing out some [partial or semipartial correlations](#)). This third variable may clarify the relationship between the first two variables or add new and unique information to our model. This third variable may also (or instead) “suck up” the information that exists in that first zero-order correlation thus making the original, direct relationship matter even less<sup>1</sup>.

<sup>1</sup>That last possibility of “sucking up” the information can be investigated by seeing if the third variable is acting as a [mediator or moderator](#).

For example, imagine we want to **predict** what makes adolescents develop an important set of cognitive skills (those executive functions, EFs). To do this, we start with an outcome—a criterion—of interest, e.g., how much a general, self-report measure of all EFs changes throughout middle and high school. We want to see what is related to that measure of growth—to predict why one teen shows strong development and another in fact becomes worse.

We could look into what is related to the EF growth by looking at a bunch of correlations. Indeed, we started by doing just that: just getting a sense of what is related to what in the data set, focusing on what is related to EF growth. But then we (I) decided to up our game: We started asking more specific questions of the data; we demanded more precise answers from it at the expense of having to make more precise assumptions.

In positioning EF growth as an outcome—a criterion or DV—and any other variables as the input / predictors / IVs, we are assuming that EF growth is a *result* of the other variables. One way to think about what we're doing is this: We set EF growth down in our model, add another variable to it as a predictor, use ordinary least squares to “regress” the predictor down to a line, and then see how steep that line is. Mathematically, this is only slightly different than conducting a correlation: In a correlation, we choose a line based on the regression of both variables down to that line; in linear regression, we only regress the predictor down to that line. Although the math is slightly different, the values we derive are the same (as long as the scores are standardized).

So why go through the trouble of doing a linear regression at all if a correlation (or even a semi-partial) gets us to the same point? Because linear regressions let us ask more precise questions and thus get more sophisticated answers. We can look at specific pieces of the model, for example looking at only the effect of one predictor on the criterion isolated from any influence of other predictors. We can also look at the error term and even run tests on it (e.g., to see if the error indeed approximates the normal distribution we assume it does). Linear models thus represent a more flexible approach that can be adapted to a wider range of data—and still generate more specific answers from it. (In fact, this approach is so flexible that—conceived of as **generalized linear models**—they undergird perhaps every analysis you will ever make.)

And so in linear regressions, we can test rather specific terms to see if those terms alone are significant. Does an adolescent's 6th-grade EF score predict EF growth in subsequent years (is the intercept term significant)? Does DBT participation matter (is the  $\beta$ -weight for that term significant)? Even if DBT participation matters, does the economic hardship a student experiences moderate the effectiveness of the DBT program (is a  $\text{DBT} \times \text{economic distress}$  term significant)?

If we already know how much economic distress a student is facing, do we really need to know about the DBT program at all?

That last question is more sophisticated than it may first seem. The questions before it (about intercept, a main effect of DBT, and a  $\text{DBT} \times \text{economic distress}$  interaction) are all answerable through an ANOVA. That last one about whether DBT adds significant and significantly new information to a model that already contains economic distress requires a bit more—and a change in perspective about what we're doing in these analyses.

Another way to think about what we're doing in linear regression is this: We gather together a set of variables to act as predictors. Using some eldritch process, we produce a value for EF growth that we would expect to see based on these predictors. Then, we see how close the actual EF growth score (for that set of conditions) is to the value we predicted it would be; if our predicted score is close to the actual score, we say we can create a good model that can well explain what determines that level of EF growth we actually see.

In other words, **not only can we build and test models, we can build two different models and see which model performs (predicts) better**. This is an idea that not only permeates (more advanced) statistics, but guides much of experimental design. One way to design a study is to

create two groups—and experimental and a control—and test if having something (the treatment given to the experimental group) is better than nothing. One way of testing significance is whether we can argue we have something (can reject the null) or whether we can't argue that we have something (cannot reject the null)<sup>2</sup>

We can still test if a predictor itself is significant while we think in terms of the model: Is the overall model more significant when that predictor is added to it? The math going on under the hood is different when we think in terms of changes in the overall model, but the end result is the same whether we test the significance of that one term (as we do in an ANOVA) or whether we test the change in significance of the overall model (as we do in this new method). (Given certain assumptions & arrangements) we get the same value for a relationship between two variables when we compute a correlation coefficient as we do when we compute a  $\beta$ -weight in a simple linear regression.

Given certain assumptions & arrangements, predictors that we find are significant when testing them as terms alone will also be significant when we test them in terms of changes to model fit. But looking at analyses in terms of changes to model fit gives us more flexibility and precision. Yes, this means we also gain yet another layer of things to learn, but I am hopeful that we can learn how to compute analyses this new way, and thus have you all gain a more powerful and flexible tool to use in your burgeoning careers.

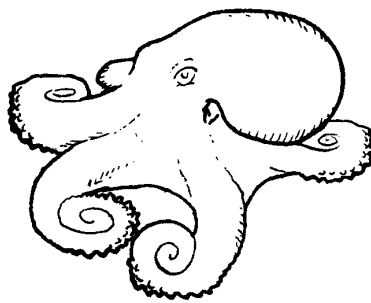
---

It may or may not be worth noting that all of the things I've discussed in this chapter are situated within the traditional statistical "*Zeitgeist*" of using models to simulate / test theories. Since the advent of "big data," however, there has been a nearly countervailing strategy of not worrying (much, if at all) about modeling relationships between variables / constructs. Instead, the focus in this other wheelhouse is making good predictions from one set of data to other sets. This other tradition often worries so little about why their analyses work that they sometimes can't even tell why it does. This underlies, e.g., the random forests and machine learning strategies that have taken firm root in a growing number of industries. And who knows? Maybe in ten years I'll be including these in a course like this. But for now, I'll only give you a few links to read more about this if somehow you're superhuman enough to have time and temerity to read them:

- Raper (2020) Leo Breiman's "Two Cultures." *Significance*, 17: 34–37. doi: [10.1111/j.1740-9713.2020.01357.x](https://doi.org/10.1111/j.1740-9713.2020.01357.x)
- Breiman (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231.
- Breiman, L, & Cutler, A. *Random Forests*.

---

<sup>2</sup>Remember, though, that a more powerful and sophisticated test is not if we have something that's better than nothing, but whether what we have is better than an alternative: which drug has fewer side effects, which type of prevention is most effective, which man is a better choice for husband.





## Chapter 5

# Variance, Covariance, Correlations, and Partial Correlations

### 5.1 The Roles of Covariance & Variance

A **standard deviation** (SD) describes how far away—on average—individual data points are from the mean. So, to compute the standard deviation, we essentially measure how far each data point is from the mean, and then take the average of those distances while standardizing this to be presented in z-scores (where one z-score equals one standard deviation). We can do this by taking each score's distance from the mean, squaring those distances, and then taking the square root of that squared distance<sup>1</sup>, and then dividing it by the number of data points (um, well, actually dividing by the number of data points *minus 1*)<sup>2</sup>. The formula for doing all of this can be written as:

---

<sup>1</sup>You'll often hear people say the reason we square the distances and then take the square root is to make all of the distances positive; otherwise, all of the distances from the mean would cancel each other out. Although this is true, couldn't we just take the absolute value of each distance instead?

Yes, we could. A major difference between taking-the-absolute-value-of-the-distances and squaring-then-square-rooting-the-distances is that squaring-then-square-rooting-the-distances amplifies the role of data points farther from the mean. And amplifying their effects was indeed one reason that method was chosen: Those who devised all of this wanted to give stronger weight to scores that are farther from the mean. Outliers and other scores that aren't as well represented by the mean were intentionally made to matter more, in part because they were considered to have more information in them than scores closer to the mean.

So yeah, it's kinda ironic that a lot of consternation is now given to removing outliers from data sets to produce more robust statistics.

<sup>2</sup>The reason we use  $(n - 1)$  instead of simply  $n$  is a good one—even if it's a rather nuanced reason. We are using a sample to estimate the values for a population, and since we don't know the population mean, we take away a degree of freedom to indicate we don't know that value. In fact, every time we have to estimate something in an analysis, we take away a degree of freedom to estimate that value. Practically, this serves to make the sample SD slightly larger than the population SD since we are dividing by a smaller number ( $1/2$  of something is larger than  $1/3$  of it). This reflects the fact that we don't know so well the actual (population) value.

We also typically use lower-case  $n$  to denote the number in a sample or subset of data set and upper-case  $N$  to denote the total number, e.g., in a population or in the whole data set, but this convention isn't always followed, so we usually just infer what the  $n/N$  refers to.

Finally, by convention, we italicize Roman letters in equations so that we know those letters denote variables; we don't italicize Greek letters, because, well, Greece is not in Italy. That, and since we don't usually write things in Greek, it's pretty clear those Greek letters denote variables. Confusingly, the way I generate the formulas in this book doesn't follow this convention and italicizes Greek letters as well; please do as I say, not as I do.

$$SD = \frac{\sqrt{\sum (X - \bar{X})^2}}{n - 1}$$

where  $X$  is the score for any given member of the sample,  $\bar{X}$  is the sample mean of that score, and  $n$  is the sample size.

Although we often report the  $SD$  in Results sections, many calculations in fact use the **variance** instead, which is nearly the same formula but without the square root<sup>3</sup>:

$$\text{Variance} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Note that I added some extra notation to the formula this time. The  $X_i$  simply means we're subtracting a particular instance of the variable  $X$ , instance  $i$ . Which is instance  $i$ ? Well, that's the point: The stuff I now put around the summation symbol ( $\Sigma$ ) explains what  $i$  means. That extra stuff around the summation symbol says that we're starting with the first instance of ( $i = 1$ , i.e., at the first data point, whatever its value) and continuing until we get the last data point (when  $i = n$ , i.e., when we've reached the last data point); in other words, "Start at the beginning

and go to the end." So, if we wrote:  $\sum_{i=3}^{n-1}$  we'd mean to start at the third instance (e.g., third row in a data matrix) and go to the next-to-last ( $n - 1$ ) instance. I point this out both so you understand it and so we can notice better what's going on when we next add another variable,  $Y$ , into the mix.

Covariance, of course, is how much two variables covary. Mathematically, we start by taking how much each pair of instances both vary from the mean:

$$\text{First variable's difference} = X_i - \bar{X}; \text{ Second variable's difference} = Y_i - \bar{Y}$$

Notice how both variables have an  $i$  as the subscript, indicating that we take the value for  $X$  for the, e.g., row as the take the value of  $Y$ . For example, for the first "instance" (e.g., row) in a data set, we say  $i = 1$  and for that row, we have  $X_1$  and  $Y_1$ ; for the second instance, we have  $X_2$  and  $Y_2$  and so on.

So, when we compute the covariance:

we first take the difference from the respective means and then multiply these differences together (before then adding these up and dividing by  $n - 1$ ), like so:

$$\text{Covariance} = \text{Cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Note three things about this:

1. Covariances can be negative (as typically conceived, variances and  $SD$ s cannot).

---

<sup>3</sup>And that's another reason we don't just take the absolute value of the distances of each point from the mean. Variance—the average *squared* distance—is used "under the hood" in calculations more often than standard deviation.

2. Covariance is maximized when  $X$  and  $Y$  are both large at the same time. Sure, that also means that there will be times when  $X$  and  $Y$  are both small at the same time, but overall, those times when both  $X$  and  $Y$  are large are what make the covariance large<sup>4</sup>. Therefore, if you want to test the true magnitude of the covariance (e.g., the correlation) between two variables, try to measure a wide range of values to allow for the possibility of times when both values are large together.
3. Because a few instances of  $X$  and  $Y$  being large together makes for a big covariance, it's vulnerable to a few outliers. Just saying.

One more thing to note about the covariance: It's represented in the same units as the variables. So, e.g., if we looked at the covariance between systolic blood pressure and lung volume, the covariance is in some weird millimeter x cubic centimeter units. This is fine—indeed necessary—if our variables are in those units, but it does often make it hard to conceive of what a covariance value *actually* means.

## 5.2 Correlations

In stats, we often standardize variables—put them all into units that are the same range—so that we can make comparisons between variables. We could do this any way we want, but commonly, we divide a value by the *SD*: This puts variables into the same range and the same units<sup>5</sup>.

We do this for covariances too, and often. Now, the covariance is the product of two variables, so we have to divide them by *both* their units, but that's easily done:

$$\text{Standardized covariance} = \frac{\text{Cov}(X, Y)}{SD_X SD_Y}$$

As you may have figured out, this is in fact the formula for Pearson's  $r$ , assuming both variables are interval / ratio.

This equation also reminds us that the concept of a signal-to-noise ratio that under-girds much of statistics is intrinsic to correlations. Here, of course, it's the covariance : variance ratio.

## 5.3 Partial & Semipartial Correlations

The concepts that underlie partial correlations are another important fundamental aspects to much of statistics—especially linear regressions. In a nutshell, a partial correlation is the correlation between two variables after first removing the effect a third variable. A semipartial correlation (confusingly sometimes called a “part correlation”) removes the effect of a third variable from just one of the two correlated variables. We often say, e.g., that we're looking at the correlation between two variables while “controlling for” the effect a third variable. (A correlation between two variables that does not account for a third (or fourth, etc.) variable is also called a zero-order or bivariate correlation.)

<sup>4</sup>This formula is the main way that outliers have an out-sized effect on what we do in frequentist statistics. They more strongly affect the variance and covariance that are in turn used in many equations.

<sup>5</sup>Yeah, we also often first subtract them from the mean, so that a 0 for one variable is comparable to a 0 from another variable.

A **partial correlation**, then, is the correlation between two variables, say  $X$  and  $Y$ , after removing the correlation each has with a third variable, say  $Z$ . Symbolizing the partial correlation between  $X$  and  $Y$  after controlling for  $Z$  as  $r_{XY.Z}$ :

$$r_{XY.Z} = \frac{\text{Correlation between } X \text{ and } Y - (\text{Correlations between } Z \text{ and both } X \text{ and } Y)}{\text{Residuals left over after removing the correlations between } Z \text{ and both } X \text{ and } Y}$$

More formally:

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

A **semipartial correlation**, remember, controls for just the correlation of one of the variable with a third variable, say only controlling for the correlation of  $X$  and  $Z$  but not for the correlation of  $Y$  and  $Z$ :

$$r_{X(Y.Z)} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)}}$$

We can create more complex partial and semipartial correlations, controlling not just for one other variable (here  $Z$ ), but for several other, additional variables.

O.K., the value of a partial correlation should be pretty clear: If I'm interested in the correlation between, say, blood pressure and breathing efficiency, I may want to control for the effects of things like age and exercise.

But why would I want to compute a *semipartial* correlation? In fact, we do this all of the time. Let me explain:

A zero-order correlation (i.e., a correlation between two variables that doesn't account for any other variables) does not make a judgment about where the variance lies—in fact, it's essentially assumed to come more or less equally from both variables. Often however, we're interested in looking at the variance in only one measure—the outcome measure. We want to know from whence comes the variation in our outcome; we are not interested in (or we try to control) the variance in our predictors. In a linear regression, we typically investigate the effect of predictor  $X$  on *outcome*  $Y$  while also controlling for the effect of predictor  $Z$  on outcome  $X$ . In other words, in linear regressions with one outcome and more than one predictor, we isolate the effects of each predictor on the outcome from the effect of other predictors on the outcome while letting each predictor fully covary with the outcome (i.e., let predictor  $X$  covary with outcome  $Y$  while removing the effect of predictor  $Z$  from the relationship between  $X$  and  $Y$ —unless we add a term for the covariance of  $X$  and  $Z$  as an interaction term).

## 5.4 Investigating Why DBT Works or Doesn't Work

Dialectical behavior therapy (DBT) is an intervention strategy similar to cognitive behavior therapy. It focuses on using mindfulness training to help regulate one's emotions and behaviors. Developed first for those with borderline personality disorders and subsequently found to be quite effective among those with suicidal tendencies, it is seen by some as holding potential use for more general populations.

Among those who believe it may help a broader range of individuals is the head of a local school. This school already has a modified—and more pervasively infused—health curriculum that seeks to help the school's adolescent students understand and control their emotions, and through that be more in control of their academic and social lives.

This school therefore implemented a DBT “curriculum” that is completed by seventh- and eighth-grade students. There are now about four years of data tracking these DBT-participating students from sixth through ninth grade. This gives us a sense of how they were before participating in the DBT program, how they did during it, and if any changes persist a while afterwards. We also have sixth- through ninth-grade data for all students from years before the school implemented the DBT curriculum; these previous years' students can serve as ersatz controls.

Initial analyses contend that the DBT program is associated with mild but significant improvements in participants' executive functions relative to this ersatz control group. DBT participation was not associated with significant changes in academic performance. Of greater concern for the school, though, was that some students benefited much more strongly from the program than other students—and that students seemed not to benefit at all.

I was recently asked by the school to lend insight into those students who did not respond well to the DBT program. The idea is to understand for whom it does and doesn't work so that the school can tailor the program to better help the students and perhaps to provide additional help for those who didn't seem to get enough from the DBT program.

I was then asked to present my findings to the school's administration. In fact, I presented it about three times, each time making my explanation (and analyses) simpler and simpler.

### 5.4.1 Your Task

How would you do if asked to do the same thing? Please therefore:

1. Use simple descriptive statistics—such as frequency counts—as well as zero-order and partial correlations
2. To investigate what factors (in the data described below) predict which students will benefit from the DBT program.
3. Please write this up in a simple 1 – 2-page report that could be understood by non-experts.

The dataset `ef_slopes.ods` is accessible in our course's BlackBoard site.

**Please note:**

*Although anonymized, these are real data. Please treat them with the respect, confidentiality, and care.*

Note as well that, as real data, there may be expected relationships between variables are not strong and unexpected relationships that are. In addition, most relationships won't be statistically significant. Please don't rely on significance as the only criterion you use to make your decisions: Instead, compare the relative magnitude of relationships to find ones you think hold more promise than other relationships; if something is indeed significant, that's nice, but we don't need to hold ourselves just to that.

### 5.4.2 Description of the Data

The variables in the dataset are:

**5.4.2.1 ID**

An identifier for each student.

**5.4.2.2 Received DBT Intervention?**

Whether the given student did (1) or did not (0) participate in the full regimen of the DBT program.

**5.4.2.3 Teacher Same Ethnicity?**

Whether the teacher is (1) or is not (0) the same ethnicity as the student.

**5.4.2.4 Teacher Same Gender?**

Whether the teacher is (1) or is not (0) the same gender as the student.

**5.4.2.5 Behavioral Regulation Slope – Teacher Report**

Each year, teachers at the school are asked to rate a subset of their students on a list of various behaviors believed to indicate a student's level of executive functioning. The scores on these behaviors are summed to create a score for each student each year. To facilitate analyses here, I computed the normalized slope for the line on which a student's yearly score regressed. **For all slopes, negative values indicate *improvements*—that a student's functioning got better.**

Behavioral regulation includes a subset of executive functions most closely related to one's overt behaviors.

**5.4.2.6 Meta-Cognitive Slope – Teacher Report**

Meta-cognition includes a subset of executive functions most closely aligned with how one thinks and feels.

**5.4.2.7 All Executive Functions Slope – Teacher Report**

All executive functions combines the scores of the behavioral regulation and meta-cognitive sub-domains into a total score.

**5.4.2.8 Behavioral Regulation Slope – Student Self-Report**

Each year, in addition to the teachers rating a subset of their students, each student also rates themselves on how they feel they've behaved vis-à-vis behaviors related to executive functioning.

The behavioral regulation slope is the regression line for the scores for each students across these four years on executive functions most related to overt behaviors. Again, *negative* slopes denote *improvements* in executive functioning.

**5.4.2.9 Meta-Cognitive Slope – Student Self-Report**

Again, these are executive functions most closely aligned with internal cognitions or emotions—here as self-reported by the students.

**5.4.2.10 All Executive Functions Slope – Student Self-Report**

The students' self-reported behavioral regulation and meta-cognitive subscores are first combined here before a slope for the regression line was computed for each student.

**5.4.2.11 Mindfulness Slope**

Recently, we added a self-reported mindfulness score by asking students to complete another instrument commonly seen to measure just that. Again, negative slopes indicate improvements.

**5.4.2.12 Emotional Regulation Slope**

The students complete a third instrument that measures their emotional regulation per se. Again, negative slopes indicate improvements.

**5.4.2.13 Discipline Incidents Slope**

Records are kept for whenever something happens on campus that leads to a student being “written up” for something that requires that student being somehow disciplined. Therefore, each student has a number of times that they are “written up” each year. This number could range from zero (for never having been written up for something requiring discipline) to sometimes rather large values.

This slope is the linear regression of this number of discipline incidents each year. Positive values therefore indicate growths in the number of times that student “got into trouble.”

This is also among the main outcome variables the school was interested in: They wanted to understand who had more discipline incidents, who subsequently got better, and—especially—whom benefited *least* from the DBT program so that these student (and those like them) could be targeted for more help.

**5.4.2.14 Self-Reflection Slope**

The linear regression for the number of times a student was put in what is essentially a mild form of in-school suspension. Positive values denote more incidents.

**5.4.2.15 In-School Suspension Slope**

The linear regression for the number of times a student was put in what is essentially *actual* in-school suspension. Positive values denote more incidents.

**5.4.2.16 Out-of-School Suspension Slope**

The linear regression for the number of times a student was suspended from school; this is the most severe of the three types of suspensions. Positive values denote more incidents.

**5.4.2.17 ELA Grade**

A 4-point grade for the student on English / language arts courses for that student's ninth grade. 4 is the highest grade possible.

**5.4.2.18 Math Grade**

A 4-point grade for the student on math courses for that student's ninth grade. 4 is the highest grade possible.

**5.4.2.19 Female?**

Whether the student does (1) or does not (0) identify as female.

**5.4.2.20 Free/Reduced School Lunch Group**

Whether the student is in the economic stratum that was previously used to allow them to be eligible for free (1) or reduced (0) school lunches, or whether they were not eligible (-1) for free / reduced school lunches.

**5.4.2.21 Economic Distress?**

Like Free/Reduced School Lunch Group, this is a measure of the student's family's economic situation, simply whether a student does (1) or does not (0) classify as experiencing economic distress.

**5.4.2.22 Mother at Home?**

Whether the student self-reported that their mother does (1) or does not (0) live with them.

**5.4.2.23 Father at Home?**

Whether the student self-reported that their father does (1) or does not (0) live with them.

**5.4.2.24 Adult Brother at Home?**

Whether the student self-reported that one or more adult brothers do (1) or do not (0) live with them.

**5.4.2.25 Adult Sister at Home?**

Whether the student self-reported that one or more adult sisters do (1) or do not (o) live with them.

**5.4.2.26 Grandparent at Home?**

Whether the student self-reported that one or more grandparents do (1) or do not (o) live with them.

**5.4.2.27 Other Adult at Home?**

Whether the student self-reported that one or more “other” adults do (1) or do not (o) live with them.

**5.4.2.28 Total Number of Adults at Home**

Simply the sum of each of the other “adults at home” variables.

**5.4.2.29 Special Education Status**

Whether a student has been (1) or has not been (o) diagnosed with a condition that makes them eligible for special education services.

**5.4.2.30 Intellectual Impairment?**

Whether a student has been (1) or has not been (o) diagnosed with an intellectual disability.

**5.4.2.31 Social Emotional Impairment**

Whether a student has been (1) or has not been (o) diagnosed with a social and/or emotional disability.

**5.4.2.32 High Risk Category**

Whether a student has been (1) or has not been (o) identified as benefiting from interventions related to high-risk behaviors.

**5.4.2.33 Years at This School**

The number of years the student has been at this school.

**5.4.2.34 Number of School Absences**

The number of times in ninth grade that the student was absent—excused or not—from school.

**5.4.2.35 Ethnicity: Asian**

A dummy variable indicating whether the student identifies as Asian (1) or not (0). Note that dummy variables are a good way to handle times when participants can identify with more than one group; some participants are identified here by more than one “ethnicity” dummy variable.

**5.4.2.36 Ethnicity: American Indian**

A dummy variable indicating whether the student identifies as American Indian (1) or not (0).

**5.4.2.37 Ethnicity: Black**

A dummy variable indicating whether the student identifies as Black (1) or not (0).

**5.4.2.38 Ethnicity: Hispanic**

A dummy variable indicating whether the student identifies as Hispanic/Latin (1) or not (0).

**5.4.2.39 Ethnicity: White**

A dummy variable indicating whether the student identifies as White (1) or not (0).

**5.5 Using SPSS****5.5.1 Accessing SPSS & Data Importation**

1. Go to CUNY Virtual Desktop: <https://www.cuny.edu/about/administration/offices/cis/virtual-desktop/>
2. Click on the SPSS icon to open it. N.b.: Your session can time out and suddenly. Make sure you save your work to an actual hard drive (or USB, online account, etc.) so you don't lose it!
3. In SPSS, with Open another file... highlighted, click Open and navigate to wherever you have ef\_slopes.ods saved.
  1. The first line indeed includes names, so leave that option checked
  2. The delimiter is indeed a comma, decimals are periods, and text marked with double quotes, so leave those options chosen
4. When you finish the data importations, SPSS opens an Output window (in addition to the Data Editor window). In this window will appear both the results of any analyses and the code that SPSS used to generate those results. This is worth explaining.
  - A benefit of SPSS is that it has an efficient and intuitive GUI, but in fact that GUI is used by SPSS to generate code that it actually uses to manipulate data and run operations.

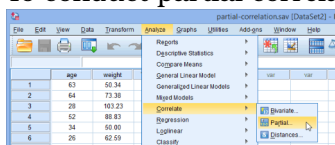
- At first, then, you can rely on the GUI to run analyses. However, as you gain experience, you can modify the code generated through the GUI before finally simply pasting / writing in code without using the GUI.
- You can also easily choose parts of SPSS's output to copy and paste elsewhere, e.g., into a manuscript (for figures, tables, etc.) or a text editor (for code).
  - For example, if we had specified formats for the variables in the Advanced Options for the data importation, we could paste out the code generated now into a file that we simply access to re-import those (or similar) data.
  - In the output, you'll notice that you can navigate from the left-hand menu, as well as delete output you no longer want. You can also right-click to rename many aspects of it and then save / export it.
- You can also access the GUI from either the Data View window or the Output window, letting you run / modify analyses either while looking at your data or the results, respectively.

5. Focusing instead on the Data Editor window, notice there are two tabs at the bottom, a Data View and a Variable View tab.

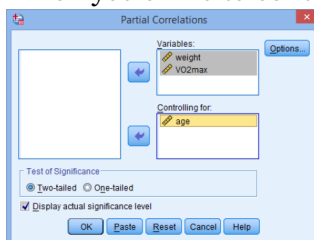
1. The Data View tab presents the data in the matrix form one is used to in a spreadsheet program.
  - Like a spreadsheet program, clicking into a cell allows one to directly modify its contents.
  - Right-clicking a column allows one to access and modify information, etc. about that variable. (Right-clicking rows has fewer options.)
    - Note that SPSS thus requires that data be formatted with variables as columns and cases as rows.

## 5.5.2 Conducting Correlations and Partial Correlations in SPSS

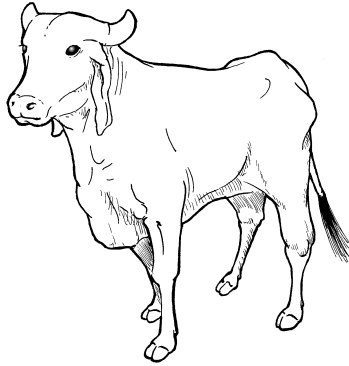
1. To conduct correlations, simply click Analyze > Correlate > Bivariate
2. In the dialogue box that opens, put the variables you wish to correlate. However, we can have SPSS compute zero-order correlations along with partial correlations.
3. To conduct partial correlations, choose Analyze > Correlate > Partial



In the Partial Correlations dialogue box, choose which variables you wish to correlate, and which you'd like to control for:



4. Click the Options button, and select, e.g., Means and standard deviations and Zero-order correlations<sup>6</sup> checkbox in the Statistics area
5. Please use these correlations along with table of, e.g., frequencies to prepare a report on insights you make into these data.



---

<sup>6</sup>Remember, zero-order correlations is the confusing term for simple, bivariate correlations between two variables.

## Chapter 6

# Effect Size: Explanation and Guidelines

Effect size is a simple idea that is finally gaining traction. It refers to a class of statistics that quantify the magnitude of a relationship or difference, independent of sample size. Most effect size statistics are standardized, so a given effect size statistic can be compared directly with that same type of effect size statistic from other analyses—or even from other studies that sample the same or similar populations.

The effect being measured can be either a difference (such as the difference between an experimental-group and a control-group mean, or the difference in number of events between groups) or an association (e.g., correlations). Different effect size statistics are computed in different ways; this means that we cannot usually directly compare one effect size statistic to another type of effect size statistic. However, the same type of effect size can be compared across different analyses or studies, and in many cases, effect size measures can be converted from one form to another (see Section 6.3).

Effect sizes are descriptive statistics. For measures of the size of an association (like a correlation), an effect size statistic may assume a linear relationship<sup>1</sup>, but they don't assume, e.g., that the population is normally distributed. Since they make few assumptions, effect size statistics are inherently robust.

Effect size statistics can complement significance tests. Significance is, of course, a yes-or-no indication of whether there is “enough” of a difference/association relative to noise: An effect is either significant or not; there are no gradations to significance. Effect size statistics do show gradations and so can be used to properly provide the nuance that people seek when they report that something is “very” or “slightly”—or even “almost”—significant. (As noted in Section 6.2 below, effect size statistics are often described as being “small,” “medium,” or “large,” but this valuation of them doesn't—well, *shouldn't*—carry anything but an arbitrary weight.)

Effect sizes can also be reported with confidence intervals, providing an informal test of significance. Since an effect size measures magnitude, while a significance test determines whether an effect is “not zero,” an effect is likely significant if its 95% confidence interval does not include zero. However, statistical significance still depends on factors such as model specification and the inclusion of covariates.

---

<sup>1</sup>In this case, it also would assume homoskedasticity. They also assume that samples are independently and identically distributed (“iid”), meaning that (a) the value of each data point in a given variable is independent from the value of all/any other data point for that variable and (b) each of those data points in that variable are drawn from the same distribution, e.g., they're all drawn from a normal distribution.

## 6.1 Common Effect Sizer Statistics

### 6.1.1 Mean Differences

These measure the distance between two or more means. Like most effect size statistics, they are also standardized (measured in terms of standard deviations) so they can be compared between studies.

#### 6.1.1.1 Cohen's $d$

One of the most commonly used effect size statistics is Cohen's  $d$ , which expresses the standardized difference between two group means:

$$\text{Cohen's } d = \frac{\text{First Mean} - \text{Second Mean}}{\text{Pooled } SD}.$$

We combine (or “pool” the  $SD$ s because there are two of them (one  $SD$  for each mean). To do this, we essentially take the average of the two  $SD$ s<sup>2</sup>.

Therefore, Cohen's  $d$  is presented in terms of standard deviations. A Cohen's  $d$  of 1 means that the means are one standard deviation apart.

You may remember that z-scores are also presented in terms of standard deviations—that a z-score of 1 means that that person's score is one standard deviation away from the mean. This isn't a coincidence and means that Cohen's  $d$  can be looked at as a z-score.

#### 6.1.1.2 Cohen's $f$ and $f^2$

Cohen introduced  $f$  as a measure of effect size for  $F$ -tests, specifically to quantify differences among three or more means. In contrast, he developed  $d$  to measure the effect size between two means. The exact formula for computing  $f$  varies slightly depending on the number of levels in the factor and the variance structure.

To extend this concept to more complex models, Cohen introduced  $f^2$ , which applies not only to ANOVA-family models but also to general(ized) linear regression. The primary distinction between  $f$  and  $f^2$  is that  $f^2$  is simply  $f$  squared. Cohen recommended using  $f^2$  for complex models because it aligns with how other parameters, such as variance-explained measures, are typically computed using squared values.

An important advantage of  $f^2$  is its flexibility: it can be used to assess the effect of a single predictor or a set of predictors, whether or not other variables in the model have been controlled for or partialled out.

More about Cohen's  $f$  can be found at this [Statistics How to](#) page.

---

<sup>2</sup>For what it's worth, we actually take the square root of the sum of the variances, and then divide that by 2, i.e.:  

$$\text{Pooled } SD = \frac{\sqrt{(SD_{\text{First Mean}}^2 + SD_{\text{Second Mean}}^2)}}{2}.$$

### 6.1.1.3 Other Measures of Mean Differences

Cohen's  $d$  is not the only measure of the effect size of mean differences—although it is the most common. Two others—Hedges'  $g$  and Glass's  $\Delta$ —are worth mentioning. All three are all standardized effect size measures used to quantify the difference between two groups in terms of standard deviations, but they differ slightly in calculation and applicability.

Table 6.1: Common Effect Size Measures of Mean Differences

Aspect	Cohen's $d$	Hedges' $g$	Glass's $\Delta$
<b>Denominator</b>	Pooled standard deviation	Pooled standard deviation with small-sample correction	Control group standard deviation
<b>Use Case</b>	Large samples, equal variances	Small samples	Unequal variances
<b>Correction Factor</b>	None	Corrects for small sample bias	None
<b>Applicability</b>	Widely used in social sciences	More accurate for small samples	Best for heteroscedastic data

#### Summary

- Use **Cohen's  $d$**  in large-sample studies with equal variances.
- Use **Hedges'  $g$**  to correct for bias in small samples.
- Use **Glass's  $\Delta$**  when group variances are expected to differ substantially.

#### 1. Cohen's $d$

- Cohen's  $d$  measures the standardized mean difference between two groups.
- **Formula:**

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}$$

where:

- $\bar{X}_1$  and  $\bar{X}_2$  are the means of the two groups.
- $s_p$  is the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- **Key Points:**

- Assumes equal variances between the groups (homoscedasticity).
- Suitable for large samples.
- Can overestimate the effect size for small sample sizes.

## 2. Hedges' $g$ (Correction for Small or Unequal Samples)

- Hedges'  $g$  is a variation of Cohen's  $d$  that corrects for the upward bias in  $d$  when sample sizes are small (usually considered when  $n < 20$ ).

- **Formula:**

$$g = d \times \left( 1 - \frac{3}{4(n_1 + n_2 - 2) - 1} \right)$$

- **Key Points:**

- Incorporates a correction factor to reduce bias in small sample sizes.
- Provides a more accurate effect size estimate when ( $n < 20$ ).
- For large samples, Hedges'  $g$  converges to Cohen's  $d$ .
- Often used in meta-analysis where comparisons between studies of very different sizes are made.

## 3. Glass's $\Delta$

- Glass's  $\Delta$  uses only the standard deviation of the control group ( $s_2$ ) as the denominator, instead of a pooled standard deviation.

- **Formula:**

$$\Delta = \frac{\bar{X}_1 - \bar{X}_2}{s_2}$$

where:

- $s_2$  is the standard deviation of the control group.

- **Key Points:**

- Useful when variances between groups are unequal (heteroscedasticity).
- May produce biased estimates if the control group standard deviation is not representative.
- Often applied in scenarios where the experimental treatment group might naturally have a higher variance (e.g., due to a treatment effect).

### 6.1.2 Proportions of Variance Explained

Cohen's  $d$  and  $f$  measure the (standardized) difference between means. Cohen's  $d$  measures it for two means, while Cohen's  $f$  is used to measure it between three or more means. Both of these statistics can be as small as zero (when there is no difference) to positive infinity. Both simply represent the number of standard deviations between the means, and if the effect size is more than 1 SD, then the effect size will be greater than 1.

Another set of effect size measures are standardized differently: They measure proportions, and so can only range between 0 and 1. The ones described in this section measure the proportion of total variance explained by a particular term in a regression model.

### 6.1.2.1 (Squared) Correlations

Perhaps the simplest measure of proportion of variance explained is correlations, specifically squared correlations. Squared correlations are indeed effect size statistics, and they measure the amount of variance explained in each of the two variables that is explained by their relationship compared to all of the variance in each of them.

For example, if the correlation between two variables is .50, i.e., if  $r = .50$ , then  $r^2 = .50^2 = .25$ . In that case, the correlation accounts for 25% of the variance in each of the variables.

### 6.1.2.2 Eta-squared ( $\eta^2$ ) and Partial $\eta^2$

The other three “proportion of variance explained” statistics are used to measure the effect size of individual terms in a linear regression model.

The first of these is **eta-squared ( $\eta^2$ )**, which quantifies the proportion of total variance in the outcome variable that is explained by a given predictor. It is calculated as:

$$\eta^2 = \frac{SS_{\text{Effect}}}{SS_{\text{Total}}}$$

This makes  $\eta^2$  conceptually similar to  $R^2$ , which measures the total proportion of variance explained by all predictors in a regression model. Like the correlation coefficient  $r$ , eta ( $\eta$ ) itself can be understood as the proportion of standard deviation differences in the outcome explained by the predictor, while  $\eta^2$  represents variance explained as a proportion of total variance.

However,  $\eta^2$  has a notable limitation: it does not account for other predictors in the model. As additional terms are introduced, the individual  $\eta^2$  values for each predictor tend to decrease, since they represent only the variance uniquely attributable to each predictor relative to total variance.

To address this, researchers use **partial eta-squared ( $\eta_p^2$ )**, which represents the proportion of variance explained by a specific predictor after accounting for other predictors in the model. Partial  $\eta^2$  is conceptually similar to partial  $r^2$ , as it isolates the unique contribution of a predictor while removing variance shared with other terms.

In a **one-way ANOVA** (i.e., a model with a single categorical predictor),  $\eta^2$  is equivalent to the overall model  $R^2$ . However, in models with more than one predictor, partial  $\eta^2$  is preferred and the overall  $R^2$  will be different than each of the partial  $\eta^2$ s.

### $\eta^2$ Compared to Cohen's $f$ and $f^2$

Cohen's  $f$  and  $f^2$  serve a similar purpose but differ in how they handle variance:

- **$\eta^2$  vs.  $f$  (ANOVA):** While  $\eta^2$  measures the proportion of variance explained by a factor,  $f$  adjusts for unexplained variance, making it more suitable for cross-study comparisons. The relationship between them is:

$$f = \sqrt{\frac{\eta^2}{1 - \eta^2}}$$

- **Partial  $\eta^2$  vs.  $f^2$  (Regression):** Partial  $\eta^2$  describes the proportion of variance explained by a predictor after controlling for other variables, while Cohen's  $f^2$  expresses the incremental contribution of a predictor relative to the unexplained variance:

$$f^2 = \frac{R^2}{1 - R^2}$$

Since  $f^2$  explicitly models the variance explained relative to unexplained variance, it is commonly used in multiple regression, particularly for power analysis and comparing models across studies.

Thus, while  $\eta^2$  and partial  $\eta^2$  are useful for describing within-sample variance explained,  $f$  and  $f^2$  provide standardized effect size measures better suited for meta-analysis and statistical power estimation.

Table 6.2: When to Use  $\eta^2$ ,  $f$ , and  $f^2$

Criterion	$\eta^2$	$f$ (ANOVA)	$f^2$ (Regression)
Use Case	ANOVA (variance explained)	ANOVA (standardized effect size)	Regression (incremental variance explained)
Interpretation	Proportion of total variance explained	Standardized measure of effect size	Standardized measure of predictor impact
Best for Comparing Studies?	No	Yes	Yes
Used in Power Analysis?	No	Yes	Yes
Inflation in Small Samples?	Yes	No	No

#### Therefore:

- Use  $\eta^2$  to describe the proportion of variance explained in ANOVA and regression models.
- Use Cohen's  $f$  for standardizing effect sizes in ANOVA, making them comparable across studies.
- Use Cohen's  $f^2$  in regression to assess the impact of specific predictors, particularly when measuring incremental effects.
- For a single dichotomous predictor, Cohen's  $d$  and  $\eta^2$  can be converted into each other, but for more complex models, additional transformations are required.

This [Analysis Factor](#) post gives a good further explanation of  $\eta^2$ . Recommendations on interpreting and reporting  $\eta^2$  are given well in [this StackExchange Q&A](#).

#### 6.1.2.3 Omega-squared ( $\omega^2$ )

$\omega^2$  is very similar to  $\eta^2$ . They both measure proportion of total variance accounted for by a given term in a model, but compute it in slightly different ways<sup>3</sup>. The way  $\eta^2$  computes it makes

<sup>3</sup>If you're curious about how the three measures— $\eta^2$ ;  $\omega^2$ ; and the next one,  $\varepsilon^2$ —are computed (from Maxwell, Camp, & Arvey, 1981, cited in Okada, 2013):

$$\eta^2 = \frac{SS_b}{SS_t}$$

it systematically overestimate the size of an effect—when it is used to measure the size of the effect for the population (i.e., when inferring from the sample to the population). Although this overestimation gets smaller as the sample gets larger, it always present (until the sample is the same size as the population).

The way  $\omega^2$ —and partial  $\omega^2$ —estimate unexplained variance makes them always smaller than  $\eta^2$  (and partial  $\eta^2$ ).  $\omega^2$  is therefore a more conservative estimate of effect size than  $\eta^2$ . Given this, many prefer  $\omega^2$  over  $\eta^2$ .

#### 6.1.2.4 Epsilon-squared ( $\epsilon^2$ )

The third and final member of our Greek-alphabet soup of stats to measure the proportion of variance explained is  $\epsilon^2$ . Everyone agrees that  $\eta^2$  overestimates the effect. Some, like Okada (2013), argue that  $\omega^2$  is sometimes *too* conservative, underestimating the true size of an effect.

$\epsilon^2$  (and partial  $\epsilon^2$ ) may be closer to “just right,” giving what may be the least biased estimate. Anyway, its value is always between the other two (or equal to them).

It’s worth noting that in a one-way ANOVA,  $\epsilon^2$  is equal to the *adjusted*  $R^2$ .

### 6.1.3 Odds & Risk Ratios

Odds ratios (ORs) and risk ratios (RRs) are often treated as standardized measures of effect size. Under appropriate conditions—i.e., comparable outcome definitions and baseline rates—they can be used to compare the magnitude of associations across studies.

Risk is simply another term for probability, and risk ratios represent the relative likelihood of an event between two groups. Both risks and risk ratios range from 0 to 1, much like proportion-of-variance metrics such as  $\eta^2$  or  $R^2$  Section 6.1.2.

In contrast, odds and odds ratios are unbounded above and can exceed 1. This asymmetry may make them less intuitive for some audiences, especially when comparing across studies. Nonetheless, it is statistically valid to compare odds or odds ratios across studies—though in some contexts, interpretability may be improved by transforming them to effect size statistics bounded between 0 and 1.

Two classic measures that do just this are the  $\phi$  (phi) coefficient and Yule’s  $Q$ . Both are designed to quantify the strength of association between two binary variables—for example, the relationship between disease status (present/absent) and group membership (exposed/unexposed). When variables have more than two categories, related measures such as Cramér’s  $V$  are more appropriate.

The  $\phi$  coefficient is defined as:

$$\phi = \frac{AD - BC}{\sqrt{(A + B)(A + C)(D + B)(D + C)}}$$

and

$$\omega^2 = \frac{SS_b - df_b MS_w}{SS_t + SS_w}$$

$$\epsilon^2 = \frac{SS_b - df_b MS_w}{SS_t}$$

where  $SS_b$  is the sum of squares between groups,  $df_b$  is the degrees of freedom between groups,  $SS_w$  is the sum of squares within each group,  $MS_w$  is mean sum of squares between groups, and  $SS_t$  is the total sum of squares (i.e.,  $SS_t = SS_b + SS_w$ ).

where  $A$ ,  $B$ ,  $C$ , and  $D$  refer to the cell counts of a  $2 \times 2$  contingency table:

	Present	Not Present
Group 1	$A$	$B$
Group 2	$C$	$D$

Despite its structural differences from the Pearson correlation coefficient,  $\phi$  is mathematically equivalent to  $r$  when both variables are dichotomous. It is also frequently used as an effect size accompanying  $\chi^2$  tests, and can be computed directly as  $\phi = \sqrt{\chi^2/n}$ .

While  $\phi$  is a valid and interpretable measure of association, it has notable limitations. It is sensitive to rare outcomes and can be inflated when marginal frequencies are highly unbalanced. This makes  $\phi$  less suitable for studies involving rare events—such as mortality rates—where other statistics may provide more stable estimates.

Yule's  $Q$  was developed to address these limitations. It is specifically designed to measure association between **odds** and is effectively a transformation of the odds ratio into a scale ranging from  $-1$  to  $+1$ , similar to correlations. Given a  $2 \times 2$  contingency table, it is defined as:

$$Q = \frac{AD - BC}{AD + BC}$$

Alternatively, it can be expressed directly in terms of the odds ratio:

$$Q = \frac{OR - 1}{OR + 1}$$

This transformation offers a symmetric, bounded, and more interpretable summary of the magnitude of association when using odds ratios.

## 6.2 “Small,” “Medium,” & “Large” Effects

Like much of statistics, Cohen's  $d$  is standardized into  $z$ -scores/ $SD$ s (remember, the formula for it is to divide it by  $SD$ s). However, simply reporting Cohen's  $d$  without interpreting what that means has a couple of disadvantages: (a)  $z$ -scores are not intuitive for lay audiences, and (b) there are other measures of effect size than Cohen's  $d$ —and they aren't all measured on the same scale. Given both of these factors, in his seminal book, *Statistical Power Analysis for the Behavioral Sciences*, Jacob Cohen (1988) gave recommendations for how to interpret the magnitude of various effect size statistics in terms of “small,” “medium,” and “large” effects.

These “criteria” for evaluating the magnitude of an effect size have become quite popular. Indeed, the adoption of effect size statistics seems to be regulated by people's uses and understandings of them in relation to these criteria. They therefore deserve further consideration.

### 6.2.1 Effect Size Criteria as Percent of Total Variance

Cohen generally defined effect sizes based on the percent of the total variance that effect accounted for<sup>4</sup>:

<sup>4</sup>These percents of variance accounted for are for zero-order correlations (i.e., correlations between two variables). The percent accounted for considered “small,” “medium,” and “large” for model  $R^2$ s are slightly higher (2%, 13%, and 26%, respectively).

- “**small**” effects account for **1%**,
- “**medium**” effects account for **10%**, and
- “**large**” effects account for **25%**.

I say that he *generally* defined them as such because he didn’t see a need to be bound to this definition, in part because he repeatedly noted—as do I here—that these criteria were arbitrary. He defined them based on percent of total variance for  $d$  and then chose “small,” “medium,” and “large” values for other effect size statistics that corresponded to those values for  $d$ .

This meant, for example, that he chose levels for correlations that don’t always match up to what one would expect by squaring the correlations to get the percents of total variances. In other words, his criteria for correlations weren’t that a “small” correlations would be  $r = .1$  (i.e., where  $r^2 = .01$ ), “medium” would be  $r = .5$ , and “large”  $r \approx .63$ . In justifying this, [he notes](#) that he is not positing these criteria levels based on strict mathematical equivalences but instead on a concerted attempt to equate the sorts of effects one would obtain with one analytic strategy with an other analytic strategy; for example, the types of effects sizes (experimental psychologists) obtain with  $t$ -tests with those they would obtain through correlations.

### 6.2.2 Effect Size Criteria as Noticeability of Effects

Although Cohen was thorough in his descriptions of these effect size criteria in terms of proportions of total variance, he was also careful to couch them in practical and experimental terms.

A “**small**” effect is the sort he suggested one would expect to find in the early stages of a line of research when researchers have not yet determined the best ways to manipulate/intervene and when much of the noise had not yet been controlled.

A “small” effect can also be considered to be a subtle but non-negligible effect: the sorts of effects that are often found to be significant in field-based studies with typical samples and manipulations/interventions. Examples Cohen gives include:

- The mean difference in IQs between twin and non-twin siblings<sup>5</sup>,
- The difference in visual IQs of adult men and women, &
- The difference in heights between 15- and 16-YO girls.

A “**medium**” is one large enough to see with the naked eye. Example Cohen gives include:

- The mean difference in IQs between members of professional and managerial occupations,
- The mean difference in IQs between “clerical” and “semiskilled” workers, &
- The difference in heights between 14- and 18-YO girls.

A “**large**” effect is one that is near the upper limit of effects attained in experimental psychological studies. So yes, the generalization of this criterion to other areas of science—including nursing research—is certainly not directly supported by Cohen himself.

Examples include:

- The mean difference in IQs between college freshmen and those who’ve earned Ph.D.s<sup>6</sup>,

<sup>5</sup>The source for this—Husén, T. (1959). *Psychological twin research: A methodological study*. Stockholm: Almqvist & Wiksell—was too old for me to see if he means mono- or dizygotic twins. But I tried!

<sup>6</sup>So, I guess a full higher education career does have a large effect on a person. And, yeah, Cohen does seem a little pre-occupied with IQ, doesn’t he?

- The mean difference in IQs between those who graduate college and those who have a 50% chance of graduating high school, &
- The difference in heights between 13- and 18-YO girls, &
- The typical correlation between high school GPAs and scores on standardized exams like the [ACT](#).

### 6.2.3 Effect Size Criteria for Odds Ratios

Cohen (1988) discussed proportions (aka risks) and presented effect size measures for a proportion's difference from .5 (Cohen's *g*) and the difference between two proportions (Cohen's *h*), which could be used to present the magnitude of a risk ratio; even though a risk ratio *per se* is already a fine effect size stat, Cohen didn't give size criteria for risk ratios, but instead for his *h*.

He didn't, however, discuss odds or odds ratios directly, and thus didn't give his opinion about what could be considered "small," "medium," and "large" values for odds or odds ratios. Yule's *Q* (Section 6.1.3) can be considered comparable to risk ratios, risk ratios weren't given size criteria either.

Chen et al. (2010) nonetheless gives some guidance by providing ranges of effect size criteria for odds ratios by comparing values with criteria for "small," "medium," and "large" Cohen's *ds*. Chen et al.'s (2010) rules of thumb for "small," "medium," and "large" odds ratios (below) deserve especial explanation.

The size of an odds ratio depends not just on the difference in outcomes in a group (e.g., the numbers of Black woman with and without pre-eclampsia), but also the difference in outcomes in a comparison group (e.g., the numbers of non-Black women with and without pre-eclampsia). It is thus not so easy to compute simple (simplistic) rules of thumbs for the sizes of odds ratios<sup>7</sup>.

In addition, the exact values for what to consider as a "small," "medium," and "large" effect depend on the overall frequency, with smaller events require larger odds ratios to equate to a given level of Cohen's *d*.

Nonetheless, Chen et al. (2010) presents some guidelines that can serve as guides in most cases. Using the median values suggested by their results:

- "Small"  $\approx 1.5$
- "Medium"  $\approx 2.75$
- "Large"  $\approx 5$

However, those suggestions can vary substantially based on the event rate in the reference group (infection rates in the non-exposed group in Chen et al.'s article):

Table 6.3: Some Suggested Odds Ratios Corresponding to "Small," "Medium," and "Large" Effect Sizes Based on the Probability of the Event in the Reference Group (from Chen et al., 2010, p. 862)

Probability of Event in Reference Group	"Small" OR	"Medium" OR	"Large" OR
.01	1.68	3.47	6.71
.05	1.53	2.74	4.72
.10	1.46	2.50	4.14

<sup>7</sup>This is also true for, e.g., risk ratios, hazard ratios, means ratios, and hierarchical models.

These estimates are based on simulations assuming a logistic model and are meant as heuristics, not rigid standards. Importantly, they illustrate that the magnitude of an odds ratio is not directly comparable across studies unless the base rates are similar.

### 6.2.4 A Few Words of Caution About Effect Size Criteria

As useful as it is to talk about effect sizes being “small” or “large,” I must underline Cohen’s own admonition (e.g., p. 42) that we use this rule of thumb about “small,” “medium,” and “large” effects cautiously<sup>8</sup>. He notes, for example, that

when we consider  $r = .50$  a large [effect size], the implication that .25 of the variance is accounted for is a large proportion [of the total variance] must be understood *relatively*, not absolutely.

The question, “relative to what?” is not answerable concretely. The frame of reference is the writer’s [i.e., Cohen’s own] subjective average of [proportions of variance] from his reading of the research literature in behavioral science. (pp. 78 – 79)

Many people—including reviewers of manuscripts and grant proposals—take them to be nearly as canonical as  $p < .05$  for something being “significant.” This is a real shame since effect sizes offer us the opportunity to finally move beyond making important decisions based on simplistic, one-size-fits-all rules.

Therefore, effect size measures, including Cohen’s  $d$ , are best used objectively to compare effects between studies—not to establish some standardized gauge of the absolute value of an intervention. This is indeed part of what is done in meta-analyses.

It is also what I suggest doing within your own realm of research: Just like Cohen himself did, review what appears to be generally agreed on as “small,” “medium,” and “large” effects within *your* research realm. These could, for example, correspond to levels of clinical significance<sup>9</sup>. Unfortunately, though, Cohen’s suggestions for *his* realm of research have become themselves canonized as the criteria for most lines of research in the health and social sciences.

Indeed, interventions and factors that have “small” effects can be quite important. This seems especially true for long-term changes, such as those one strives for in educational interventions or for the [subtle but persistent effects of racism](#). Teaching a diabetic patient how to check their blood insulin may have only a small effect on their A1C levels in a given day, but can save their life (or at least a few toes) in the long run.

Given this, Kraft (2020) used a review of educational research to [suggest](#) different criteria for gauging what should be considered as “small,” “medium,” or “large” effects in education interventions. His recommendations are also presented below.

<sup>8</sup>Cohen also only directly considered these criteria as they applied to experimental psychology—not, e.g., the health sciences. Indeed, he [elsewhere](#) notes that what experimental psychologists would call a “large” effect would be paltry in the physical sciences.

<sup>9</sup>With, say, the target level of outcome denoting a “medium” effect. Reaching  $\frac{1}{3}$  of that target could denote a “small” effect, and reaching  $\frac{2}{3}$ s more (167%) a “large” one. (This corresponds to the range between many of Cohen’s criteria. For example, criteria for  $r$  are .1, .3, and .5.

### 6.2.5 Table of Effect Size Statistics

Table 6.4: Effect Size Interpretations

Statistic	Explanation	Small	Medium	Large	Reference
<b><i>d</i></b>	Difference between two means	0.2	0.5	0.8	Cohen (1988, p. 25)
<b><i>d</i></b>	For education interventions	0.05	< .2	≥ .2	Kraft (2020)
<b><i>g</i></b>	Hedges' modification of Cohen's <i>d</i> for small samples	0.2	0.5	0.8	Hedges (1981)
<b><i>h</i></b>	Difference between proportions	0.2	0.5	0.8	Cohen (1988, p. 184)
<b><i>w</i></b> (also called <b><i>φ</i></b> )	$\chi^2$ goodness of fit & contingency tables. $\phi$ is also a measure of correlation in $2 \times 2$ contingency tables, and ranges between 0 and 1.	0.1	0.3	0.5	Cohen (1988, p. 227)
<b>Cramer's <i>V</i></b>	Similar to $\phi$ , Cramer's <i>V</i> is used to measure the differences in larger contingency tables. Like $\phi$ (and other correlations) it ranges between 0 and 1.	0.1	0.3	0.5	Cohen (1988, p. 223)
<b><i>r</i></b>	Correlation coefficient (difference from $r = 0$ )	0.1	0.3	0.5	Cohen (1988, p. 83)
<b><i>q</i></b>	Difference between correlations	0.1	0.3	0.5	Cohen (1988, p. 115)
<b><math>\eta^2</math></b>	Parameter in a linear regression & AN(C)OVA	0.01	0.06	≥ .14	
<b><i>f</i></b>	AN(C)OVA model effect; equivalent to $\sqrt{f^2}$	0.1	0.25	0.4	Cohen (1988, p. 285)
<b><i>f</i></b>	For education interventions (i.e., <i>f</i> equivalent for Cohen's <i>ds</i> suggested by Kraft.)	0.025	< .1	≥ .1	Kraft (2020)
<b><i>f</i><sup>2</sup></b>	A translation of $R^2$	0.02	0.15	0.35	<ul style="list-style-type: none"> <li>• For multiple regression / multiple correlation, Cohen (1988, p. 413);</li> <li>• For multivariate linear regression, multivariate <math>R^2</math>, Cohen (1988, p. 477)</li> </ul>

Statistic	Explanation	Small	Medium	Large	Reference
<b>OR</b>	Odds ratio; can be used as effect size for Fisher's exact test and contingency tables in general.	1.5 (or 0.67)	2.75 (or 0.36)	5 (or 0.20)	Chen et al. (2010, p. 862)

## 6.3 Converting Between Effect Size Measures

Most effect size statistics can be converted into other ones, but the process isn't always possible or direct (or requires additional assumptions). Table 6.6 presents the effect sizes statistics covered here that *can* be converted (and the conditions/assumptions required for that); Table 6.7 presents the effect size statistics that can't be meaningfully converted.

More usefully, Table 6.5 presents the formulas for convert between the effect size statistics that can be readily & meaningfully done.

Perhaps even more usefully, this [handy Excel spreadsheet](#) can convert between Cohen's  $d$ ,  $r$ ,  $\eta^2$ , odds ratios, and area under the curve.

In [Chapter 7](#) of their book on meta-analysis, Borenstein et al. (2011) also cover well the conversions between measures. Finally, the [effectsize](#) package for R can both compute and convert between many effect size measures, including all those mentioned here.

Table 6.5: Formulas to Convert Between Common Effect Size Statistics

From ↓ To →	Cohen's $d$	Hedges' $g^{IO}$	Pearson's $r$	$\eta^2$	$f$	$f^2$	$\phi$ , $V$ (2×2 only)	OR (logistic approx.)
$d$	–	$g = d \cdot \left(1 - \frac{3}{4N-9}\right)$	$r = \frac{d}{\sqrt{d^2+4}}$	$\eta^2 = \frac{d^2}{d^2+4}$	$f = \frac{d}{2}$	$f^2 = \frac{d^2}{4}$	$\phi = \frac{d}{\sqrt{d^2+4}}$	$\frac{d}{\pi} = \frac{\ln(OR) \cdot \sqrt{3}}{\pi}$
$g$	$d = \frac{g}{1 - \frac{3}{4N-9}}$	as $d$	as $d$	as $d$	as $d$	as $d$	as $d$	as $d$
$r$	$d = \frac{2r}{\sqrt{1-r^2}}$	as $d$	–	$\eta^2 = r^2$	$f = \frac{r}{\sqrt{1-r^2}}$	$f^2 = \frac{r^2}{1-r^2}$	$\phi = r$	–
$\eta^2$	$d = \sqrt{\frac{4\eta^2}{1-\eta^2}}$	as $d$	$r = \sqrt{\eta^2}$	–	$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$	$f^2 = \frac{\eta^2}{1-\eta^2}$	–	–
$f$	$d = 2f$	as $d$	$r = \frac{f}{\sqrt{f^2+1}}$	$\eta^2 = \frac{f^2}{1+f^2}$	–	$(f)^2 = f^2$	–	–
$f^2$	$d = 2\sqrt{f^2}$	as $d$	$r = \sqrt{\frac{f^2}{1+f^2}}$	$\eta^2 = \frac{f^2}{1+f^2}$	$f = \sqrt{f^2}$	–	–	–
$\phi$ or $V$	$d = \frac{2\phi}{\sqrt{1-\phi^2}}$	as $d$	$r = \phi$	$\eta^2 = \phi^2$	–	–	–	–
OR	$d = \frac{\ln(OR) \cdot \sqrt{3}}{\pi}$	as $d$	–	–	–	–	–	–

Table 6.6: Common Effect Size Statistics That Can Be Converted into Each Other

This Effect Size Statistic...	Can Be Converted To...	Under These Conditions
Cohen's $d$	$g, r, \eta^2, f, f^2, \text{OR}, \phi$	Assumes continuous, normally distributed data; OR/ $\phi$ require dichotomous approximation
Hedges' $g$	$d$	Hedges' $g$ is a modification of Cohen's $d$ for small sample sizes
Pearson's $r$	$d, f, f^2, \eta^2$	Assumes linear relationship
$\eta^2$	$r, f, f^2, d$	Limited to ANOVA models
Cohen's $f$	$d, r, \eta^2, f^2$	In ANOVA models
Cohen's $f^2$	$R^2, f, r, \eta^2$	In multiple regression contexts
Cohen's $w$	$\phi$ or $V$	In $2 \times 2$ tables
Cramér's $V$	$\phi, w$	Only for $2 \times 2$ ; not convertible to $d, f$ , etc.
$\phi$	$r, w, V, d$ (with assumptions)	In $2 \times 2$ tables; n.b., this is an approximate $d$ conversion
Odds Ratio (OR)	$d$ (approx.), log-OR	Approximate only; assumes logistic distribution
Risk Ratio (RR)	$d$ (approx.)	Approximate only; assumes log-binomial model

Table 6.7: Common Effect Size Statistics That **Cannot** Be Converted into Each Other

Pair	Why Not Convertible
$h \leftrightarrow d, f, r$	$h$ is based on arc-sine transformed proportions (i.e., a different metric)
$q \leftrightarrow d, f, r$	$q$ compares correlations (via Fisher's $z$ )
$V \leftrightarrow f$	$V$ is for categorical data (chi-square); $f$ for continuous
$\text{OR} \leftrightarrow r, f$ (directly)	Only approximate; depends on baseline prevalence
$\text{RR} \leftrightarrow$ anything else (except OR)	RR has no meaningful transformation outside risk models

### 6.3.1 A Few Notes on Conversions

In addition to simply listing the formulas for possible conversions, there are a few more points to make—and a couple more conversions that are worth knowing. Below are further considerations about converting Cohen's  $f$  (and  $f^2$ ) to Cohen's  $d$  and about converting relevant effect size stats into the  $t$ -scores and  $F$ -scores used to test mean differences.

### 6.3.2 Cohen's $f$ (and $f^2$ ) to Cohen's $d$

Cohen's  $f^2$  (and  $f$ ) measures the effect size of an entire model (usually an ANOVA). Cohen's  $d$  measures the effect size between two levels of single variable<sup>10</sup>. So, in order to convert between

<sup>10</sup>Hedges'  $g$  can be converted to any other effect size that Cohen's  $d$  can be converted. To convert to Hedges'  $g$  instead of  $d$ , multiply  $d$  in the given equation by  $(1 - \frac{3}{4N-9})$ .

<sup>11</sup>Remember, Cohen's  $d$  is just the difference between two means that is then standardized.

$f^2$  and  $d$ , we have to know more about the model. For a one-way ANOVA with two groups<sup>12</sup>,  $d = 2f = 2\sqrt{f^2}$ . In this particular case, then,  $f = \frac{d}{2}$ .

More generally, when there is only one term in the model:

$$f^2 = \frac{d^2}{2k}$$

It gets a bit more complicated when there are more than one terms in the model. [This site](#) covers some common situations.

### 6.3.3 Cohen's $d$ and Student's $t$

This is the  $t$  in  $t$ -test. The only additional piece of information we need to know to transform between Cohen's  $d$  and Student's  $t$  is the sample size,  $N$ :

$$t = d \times \sqrt{N}$$

$$\text{Cohen's } d = \frac{t}{\sqrt{N}}$$

### 6.3.4 $\eta^2$ and $F$ -scores

This  $F$ -test score that is used in ANOVA-family models. Like the relationship between  $d$  and  $t$ , the only additional things we need to know to compute  $\eta^2$  from  $F$  are degrees of freedom (which are closely related to sample size). Here, though, we have degrees of freedom in both the numerator (top) and denominator (bottom<sup>13</sup>):

$$\eta^2 = \frac{F \times df_{Effect}}{F \times (df_{Effect} + df_{Error})}$$

So,  $\eta^2$  is dependent on the ratio of the  $df$ s allotted to the given effect and the  $df$ s allotted to it's corresponding error term. Since we have the effect's  $df$ s in both the numerator and denominator, their effect will generally cancel out; this suggests that having more levels to a variable doesn't appreciably affect the size of its effect. However, being able to allot more  $df$ s to error does help us see the size of whatever effect is there. Larger samples won't really change the size of the effects we're measuring, but they can help us see ones that are there.

## 6.4 Additional Resources

- [Psychometrica](#) offers a wonderful and pretty thorough list of effect size measures along with freeware apps to compute them at [https://www.psychometrica.de/effect\\_size.html](https://www.psychometrica.de/effect_size.html)

<sup>12</sup>Which is itself really just a  $t$ -test but using an ANOVA framework instead

<sup>13</sup>My mnemonic to remember which is which is to think of the saying, "The lowest common denominator."

- Hojat, M. & Xu, G. (2004). A visitor's guide to effect sizes: statistical significance versus practical (clinical) importance of research findings. *Advances in Health Sciences Education: Theory and Practice*, 9(3), 241–249. doi: 10.1023/B:AHSE.0000038173.00909.f6
- Reichel, C. (2019). Statistics for journalists: Understanding what effect size means. *The Journalist's Resource*.
- [Psychometrica.de](https://www.psychometrica.de/), this very useful site contains:
  - Easy functions to compute every, commonly-used effect size measure
  - Convert between  $d$ ,  $r$ ,  $f$ ,  $OR$ ,  $\eta^2$ , and common language effect size statistics
  - Table of “small,” “medium,” and “large” effects laid out and interpreted somewhat differently than I did here
  - List of relevant sources
- [FasterCapital's Phi Coefficient: The Phi Coefficient and Yule s Q: Pioneers in Measuring Association](https://www.fastercapital.com/entry/Phi-Coefficient-and-Yule-s-Q-Pioneers-in-Measuring-Association.php) provides a very readable and thorough coverage of those two statistics.

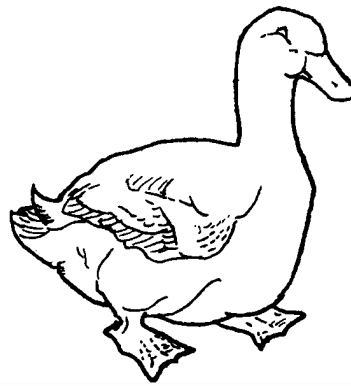


Figure 6.1: Cohen's duck

# Chapter 7

## Missing Data

*Only the beauty of what is missing can endure.* – Source unknown

***This chapter is an other “stub” that I am slowly working on. The few things I’ve added, though, seemed worthwhile enough to make this one public.***

### 7.1 Missing Data: More Than Just a Smaller Sample

#### 7.1.1 Sources of Attrition

The difference between MAR and MCAR is not trivial. Data may be missing for real and even evident reasons, but these reasons may not necessarily affect the conclusions we draw from them. In a review of attrition among new mothers participating in a 15-year study, Gustavson et al. (2012) found that education levels were lower amount those who dropped out, suggesting important sources of bias for generalizations of their results to other populations. However, those who dropped out did not differ in terms of the psychological and interpersonal/social factors central to their study. They also found that attrition did not affect the associations between those central variables. Their results suggest that even systematic biases in attrition may not significantly affect our conclusions—even if this is normally impossible to test and confirm.

Citing Davis et al. (2002), Teague et al. (2018, p. 1) similarly noted that “[s]ystematic attrition in longitudinal research occurs most often in older, non-white male participants with limited education and/or multiple health problems. Long duration and repeated assessments can also increase attrition due to the significant burden on participants.” Nonetheless, it may be useful to consider more broadly the factors found to be associated with attrition. Gustavson (2012, p. 2) provides a nice review of such factors within field-based health research:

Socio-demographic variables, such as low educational level, being out of work, and not being married, are typically related to increased risk of non-response and attrition in epidemiological studies [2,4,5,8-12]. In addition, unhealthy life style factors, such as smoking, high alcohol consumption, and physical inactivity, are related to non-participation and attrition [8,11-13].

High levels of psychological distress can predict attrition in high-risk populations, such as psychiatric outpatients and former hospitalized patients [3,14]. In populationbased

studies, psychological distress has been found to have no effect or a weak to moderate effect on attrition after adjusting for other variables [2,4,9,10]. Attrition may also be related to social factors, such as support from spouse or friends, and child's characteristics. Poor relationship quality is an important predictor of mental health problems [15]. However, social networks and support did not predict attrition in a 15-year follow-up study [5], and marital satisfaction and spousal support did not predict attrition in a job satisfaction study [6]. More knowledge is needed about the association between attrition and psychological as well as social factors.

Studies with high-risk populations found that externalizing problems and psychopathology in general among children were associated with a higher risk of parents dropping out [16,17], whereas child characteristics such as temperament, anxiety, and attention problems did not predict attrition in population-based studies [18,19]. It may be that the ways different factors affect attrition are dependent on whether the original sample was drawn from a high-risk population.

### 7.1.2 Addressing Attrition

Teague et al. (2018) conducted a meta-analysis of factors contributing to and reducing attrition in a rather wide range of longitudinal, field-based studies. They found that:

after controlling for study duration and number of waves, studies that utilised any barrier-reduction strategy had higher retention rates than those that did not use a barrier strategy (median retention using barrier strategies = 81.1%; median retention not using barrier strategies = 70.7%;  $b = 0.61$ ,  $p = .01$ ). Again after controlling for the study duration and number of waves, surprisingly, articles that reported use of at least one follow-up/reminder strategy had lower retention rates when compared to studies that did not utilise any follow-up/reminder (median retention using follow-up/reminder strategies = 76.4%; median retention not using follow-up/reminder strategies = 86.1%;  $b = -0.32$ ,  $p < .01$ ). No relationships were found between retention rate and the use of any community-building or tracing retention strategies" (p. 11)

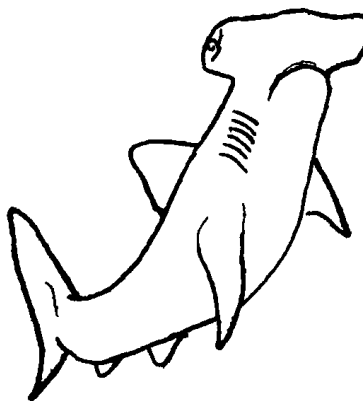
Further details about what did and did not affect retention are in this reproduced table:

- Missing data

- [Good review of caveats with Little's test in description](#)
- <https://www.sciencedirect.com/science/article/pii/S0895435618308710>
- <https://www.researchgate.net/deref/http%3A%2F%2Fwww.talkstats.com%2Fshowthread.php%2F17Multiple-Imputation-or-FIML>
- [https://www.researchgate.net/profile/Rafael-Garcia-26/post/What\\_proportion\\_of\\_missing\\_data\\_is\\_t+Revision+2013+%281%29.pdf](https://www.researchgate.net/profile/Rafael-Garcia-26/post/What_proportion_of_missing_data_is_t+Revision+2013+%281%29.pdf)
- <https://www.researchgate.net/deref/http%3A%2F%2Fwww.personal.psu.edu%2Fjxb14%2FM554%2F>
- Chapter 9 in Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modelling*. London: Sage
  - \* <https://www.stats.ox.ac.uk/~snijders/mlbook.htm>

**Table 3** Meta-analytic regression results between retention strategy themes and retention rate

	Estimate	CI (Lower - Upper)	P	I <sup>2</sup>
Model 1: Continuous total number of retention strategy types				99.86%
Barriers	0.17	0.03–0.32	0.02*	
Community	–0.03	– 0.18 - 0.11	0.63	
Follow-up/reminder	–0.15	–0.29 - -0.01	0.04*	
Tracing	0.11	–0.06 - 0.27	0.22	
Study duration	–0.04	–0.08 - 0.00	0.06	
Number of waves	0.00	–0.02 - 0.03	0.81	
Model 2: Binary usage of retention strategy types				99.84%
Barriers	0.35	–0.15 - 0.86	0.16	
Community	0.35	–0.14 - 0.83	0.16	
Follow-up/reminder	–0.83	–1.40 - -0.27	0.00**	
Tracing	0.11	–0.36 - 0.59	0.64	
Study duration	–0.03	–0.08 - 0.01	0.10	
Number of waves	0.01	–0.02 - 0.03	0.61	
Model 3: All individual strategies with p < 0.1				99.85%
Tracing - Locator form documentation	0.59	–0.44 - 1.62	0.26	
Follow-up - Reminder Phone call	–0.72	–1.20 - -0.25	0.00**	
Community - Thank you and birthday cards	0.44	–0.11 - 0.98	0.12	
Barriers - Site and home visits	0.42	–0.05 - 0.88	0.08	
Barriers - Consistency in research staff	0.39	–0.42 - 1.20	0.34	
Barriers - Alternative method of data collection	0.59	0.14–1.05	0.01**	
Study duration	–0.04	– 0.08 - – 0.00	0.05*	
Number of waves	–0.00	–0.03 - 0.02	0.89	

\* $p < .05$ \*\* $p < .01$ **Figure 7.1: Meta-Analytic Regression Results Between Retention Strategy Themes and Retention Rate**



## Chapter 8

# Linear Regression Modeling with SPSS, Part 1: Introduction

**Statistics:** *A subject which most statisticians find difficult but which many physicians are experts on.* – Stephen Senn, *Statistical Issues in Drug Development*, p. 4

### 8.1 Overview

This chapter covers the relationship between partial correlations and linear regressions before exploring and interpreting results of linear regressions conducted on the adolescent executive functioning data.

### 8.2 Core Concepts

#### 8.2.1 Linear Relationships

Very few relationships in healthcare are truly linear. There are sweet spots in how much or what sorts of care to give, sometimes diminishing returns, sometimes growing returns. This is, in fact, true for much outside of the physical sciences; even the effect of intelligence on income & wealth appears to have a non-linear ([diminshing](#)) effect.

And yet, in many cases it's good enough to assume relationships are linear. It can account for much of the relationship while being easy to model statistically. Even if we suspect that the relationships between predictors and outcome are non-linear, we often first test those relationships against a model that assumes linearity because that may well still be sufficient. In addition, seeing how well the data are fit by a linear model lets us then subsequently see *how much better* a given non-linear relationship explains it: We can even quantify and test the significance of the improvement of a non-linear model over a linear one.

### 8.2.2 Consider Removing the Intercept

O.K., removing the intercept isn't a core concept, but it can be a good idea nonetheless. One of the best predictors of the future is the past<sup>1</sup>, so simply knowing where participants are when they start participating is often among the most efficient and powerful ways of knowing where they will be later on in the study.

But wouldn't it be nice to know *what* about their initial state matters most later one? Leaving the intercept in lets it “suck up” a lot of information that could otherwise be explained by other terms in your model. Removing the intercept may allow that information to “flow back” into other predictors to allow those other predictors to explain more of your outcome.

Removing the intercept also frees up the degree of freedom used to estimate its value. This gives our other analyses a bit more power. No, that's not usually a lot, but it does help us maximize the information in our data, making us that much more efficient and conscious of the real value of data.

## 8.3 Introduction to Linear Regression Models

Correlation and linear regression are closely related tools that both describe the linear association between two variables. Correlations quantify the strength and direction of a linear relationship without making assumptions about the sources of error. A linear regression, on the other hand, models that relationship explicitly, allowing us to quantify and test different parts of that relationship. Linear regressions also allow us to model more complex relationships and—or course—to test the associations between more than two variables.

Throughout this tutorial, we will use a subset of variables from the [Add Health study](#), which measured a set of “rich demographic, social, familial, behavioral, psychosocial, cognitive, and health survey data” for over 20,000 people in five waves of data collection from 1995 to 2018. It's also worth investigating since their website now bears the [mark of MAGA](#), stating that “[o]n March 31, 2025, as a sponsor of this project, NIH requested that the following language be added to this website: This repository is under review for potential modification in compliance with Administration directives.” I've preserved some of the data [here](#); we will be using that data now with SPSS (v. 31). More about using SPSS through CUNY's Apporto service is given in Section 17.2.1.

### 8.3.1 Correlation vs. Simple Linear Regression

Let's begin by comparing the results of a simple linear regression against the results of a zero-order correlation containing the same variables. A simple linear regression is a linear regression that contains only one predictor (like a one-way ANOVA).

The Add Health data are in [long format](#), so let's first subset them to just the first wave—i.e., the first time data were collected. With the `add_health.sav` file open in SPSS:

---



<sup>1</sup>An interesting take on this—albeit one that's tangential to what we're talking about—is given in this quote from Funk's (2023) *New York Times* article:

“A world in which computers accurately collect and remember and increasingly make decisions based on every little thing you have ever done is a world in which your past is ever more determinant of your future. It's a world tailored to who you have been and not who you plan to be, one that could perpetuate the lopsided structures we have, not promote those we want. It's a world in which lenders and insurers charge you more if you're poor or Black and less if you're rich or white, and one in which advertisers and political campaigners know exactly how to press your buttons by serving ads meant just for you. It's a more perfect feedback loop, a lifelong echo chamber, a life-size version of the Facebook News Feed. And insofar as it cripples social mobility because you're stuck in your own pattern, it could further hasten the end of the American dream. What may be scariest is not when the machines are wrong about you — but when they're right.”

1. From any window, click on Data > Select Cases...
2. In the dialogue that opens, click the radio button next to If condition is satisfied and then click the If... button right below it.
3. Choose Wave from the list on the left and click the arrow button (→) to move it to the field on the right.
4. Either type = 1 or use the buttons to write that out<sup>2</sup>.
5. Click Continue and then OK.

Note that we are leaving Filter out unselected cases selected (under Options in the main dialogue). This simply flags the unwanted rows but does not otherwise change the data. Copy selected cases to a new dataset or Delete unselected cases actually changes the data. I try to leave data as intact as I can to prevent mistakes and to be able to replicate what I did.

After doing this, you will notice two things in the Data Editor window. First, in the Data View tab, rows where the Wave is *not* 1 have a line drawn through their row number:

	 AID	 Wave
1	57100270	1
2	57101310	1
<del>3</del>	57101310	2
<del>4</del>	57101310	4
<del>5</del>	57101310	5
6	57103171	1
7	57103869	1
<del>8</del>	57103869	2
<del>9</del>	57103869	4
10	57104553	1

Second, in the Variable View tab, there is now a new filter\_\$ variable added to the end of the data<sup>3</sup>. The Label for this variable summarizes what we did: Wave = 1 (FILTER). This new variable in fact defines which rows to filter out; deleting this variable will remove the filter.

<sup>2</sup>You can also simply type out Wave = 1. Choosing the variable from the list simply helps prevent typos.

<sup>3</sup>Yeah, it's also there to the far right in the Data View tab.

### 8.3.1.1 Correlation

1. With the add\_health.sav file open and subset to Wave = 1, choose Analyze from the menu bar (from any window), and click on Correlate > Bivariate<sup>4</sup>.
2. Select Participated in DBT [DBT] and Ball Executive Functions Slope – Student Self-Report [All\_EFs\_SR\_Slope] to add to the Variables field.
  1. This is a **point-biserial correlation** (i.e., a dichotomous variable correlated with a continuous), but the formula for that derives to be computationally the same as that for the Pearson, so select (or leave selected) that option under Correlation Coefficients.
  2. The default  $\alpha$ -value for rejecting the null in SPSS is .05. i.e., we are accepting a 5% chance that any given hypothesis test will be a Type 1 error—here that the correlation is equal to zero (and the Type I error being that we think it isn't equal to zero when in fact it is). We don't really have an *a priori* reason to believe that this correlation will be above or below zero, so we will divide that  $\alpha = .05$  into two pieces, letting us test if it is above zero 2.5% of the time and below zero 2.5% of the time. This is called a two-tailed test, so let's leave the Test of Significance to Two-tailed.
  3. Under the Options dialogue, we can include Means and standard deviations under Statistics. We also should leave the Missing Values option to Exclude cases pairwise. Pairwise exclusion in SPSS means that a given case (row) will be excluded from a given analysis if that row is missing data relevant to that analysis per se. Listwise exclusion in SPSS means that a row will be excluded if *any* data are missing for that case in any of the variables selected for that family of analyses—even if one values relevant to that particular analysis are both present. Listwise exclusion is nearly always too conservative a criterion, thus opening us up to biases in our analysis that come from biases introduced by variables that aren't even in that analysis.
3. In the Descriptive Statistics the Output window, we see that mean for the DBT variable is .19; since this is a dummy variable (where 1 = participating and 0 = not participating), this is also the proportion of cases that participated in the program: 19% of the students here participated in the DBT program.
4. In the Correlations table, we see that the correlation between DBT and All\_EFs\_SR\_Slope is -.165. This indicates that as we go from a DBT score of 0 to 1, the slope changes  $-.165^5$ . The Sig. (2-tailed) row in the table indicates that the  $p$ -value for that correlation is .003, which is less than the  $\alpha$ -level we established ( $\alpha = .05 / 2 = .025$  for each tail), making this correlation significant<sup>3</sup>. We could write this in a Results section as:

The point-biserial correlation between DBT participation and changes in all executive functions was significant ( $r_{pb} = -.165$ ,  $df = 326$ ,  $p = .003$ ), indicating that participating in the DBT program was associated with significantly more negative slopes (i.e., significantly greater improvements) in total executive function scores.

Which attempts to explain the relationship in simply terms that rely on little in-article jargon and acronyms. The support for this plainer-English description is supported (parenthetically) by the numerical statistics.

<sup>4</sup>Note that in SPSS the “A” in Analyze is underlined, that the “C” in Correlate is underlined, and the “B” in Bivariate is underlined. You will notice that all menu options have one letter underlined; once you are let enough to use keyboard instead of the mouse, this is the key you will type to select that option. So, to access this analysis, you could simply hold down the Alt key and type A then C then B instead of using the mouse.

<sup>5</sup>The reverse interpretation—that a 1-point decrease in slope makes it 16% more likely that that student was a member of the DBT group—is plausible for a correlation, but doesn't really make sense practically since whether a child participated depended only on the year they were admitted to the school.

### 8.3.1.2 Linear Regression

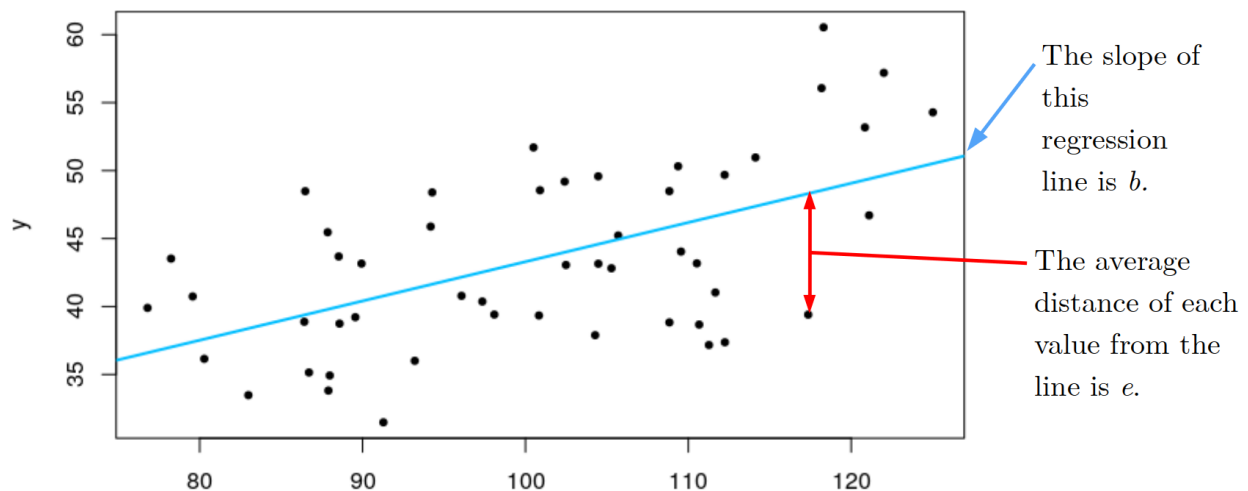
Remember that in simple correlations, we assume that unique variance / error comes equally from both variables. We formalize this mathematically by having the variance they share—their **covariance**—divided by the unshared variance from both variables:

$$\text{Correlation} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)\text{Var}(Y)}$$

Unlike a correlation, in a linear regression, unique variance / error is assumed to come from only the predictor variable(s):

$$Y = bX + e$$

where  $b$  is the slope of the regression line that best fits the cluster of  $X$  values plotted against  $Y$  and  $e$  is the average distance<sup>6</sup> of each individual  $X$  value from that line—the variance unique to  $X$  that is relegated to error:



Although our assumptions are different, we are still doing the same basic function: determining a line of best fit computed by minimizing the unique variance in our data—here in the values of our criterion,  $X$ <sup>7</sup>.

Separating out error like is done in a linear regression will eventually allow us more flexibility in how we deal with error. For now, I simply want to show the similarity between a correlation and a linear regression.

1. In whatever window you have active, click Analyze > Regression > Linear....
2. Add ZAll\_EFs\_SR\_Slope to the Dependent field and DBT to the Independent(s) field. SPSS is calling the outcome “Dependent” and the predictor(s) “Independent(s)”, as in DVs and IVs.

<sup>6</sup>In ordinary least squares, it's the square root of the mean squared distance—just as the standard deviation is square root of the deviance, which is itself the sum of squared distances from the mean.

<sup>7</sup>By minimizing the variance relegated to error, we are trying to minimize the amount of information in the data that is lost—unexplained—by the model. The better that a model is at explaining the variance—the information—in the data, then the less information is lost to error.

3. Under Statistics, have Estimates and Confidence intervals selected under Regression Coefficients (the latter since confidence intervals are slowly replacing up-and-down significance tests), and Model fit, Part and partial correlations, and—might as well—Descriptives also selected. For now, leave R squared change unselected.

(Casewise diagnostics lets you evaluate whether there are extreme outliers in the data that may be skewing the results. Durbin-Watson tests whether there is nonignorable autocorrelation between the errors of longitudinal data, with a value of “2” indicating ignorable autocorrelation and values approaching either 0 or 4 indicating that error values are not independent of each other and thus that, e.g., one should consider the nested nature of the data, q.v. Chapter 11.)

4. Under Options, select Exclude cases pairwise under Missing values for the reasons discussed above; Replace with mean is a defensible strategy for handling missing data, although multiple imputation would be preferred in all ways ... were it easy performed in SPSS.

We will not be use any Stepping Method Criteria, so the default (or really any values since this doesn't apply) are fine.

Finally, leave Include constant in equation selected. The constant of which they speak is the intercept since not all of our variables are standardized.

5. We will ignore, e.g., the Method of entering or removing terms from the model for now.

For Zscore: All Executive Functions Slope -- Student Self-Report in the Descriptive Statistics table in the Output window, we see that the Mean is .00000000 and the Std Deviation is 1.00000000, as they should be since *that* variable is indeed standardized here<sup>8</sup>.

For Participation in DBT Program?, the mean is .19 (sample size is 670); since this is a dummy variable, this means that 19% of the cases had 1s, i.e., that 19% of the students participated. The results for the Correlation also return the correlation (-.165) and that it is significant.

After the Variables Entered/Removed table (which is not relevant now), the results present a Model Summary table. This table presents statistics about how well the model overall performs when trying to fit the data we fed it. Remember how one way of thinking about linear regression models is that they try to minimize unexplained variance in the data—that they try to account for as much of the variance / information in the data as possible. The R and R Square values in this table do just that; they describe how much variance in the data set are accounted for by this particular model (containing—for now—simply whether the student participated in the DBT program). We see from this table that the R-value is .165. This, of course, is the same absolute value as the correlation between DBT and All\_EFs\_SR\_Slope. A linear regression will generate the same (or nearly same) values as a correlation on those same variables—the difference, though, is in the assumptions we're making about the data: Correlations assume error is shared equally whereas linear regressions separate out error and explicitly model it as a term among the predictors.

The R Square value in that table is just the R value squared. Squaring a correlation coefficient (i.e., computing  $r^2$  from  $r$ ) computes the shared variance between the two variables. Similarly, R Square (i.e.,  $R^2$ ) computes the variance within that data set that is accounted for by this model. Capital  $R^2$  is used to denote the variance in the data accounted for by *all* of the model terms—intercept, main effects, interactions, etc., but not error. Lower-case  $r^2$  is used to denote the variance shared by just two variables—not the whole model. (The Adjusted R Square reduces the value a bit for each term added to the model since one can improve the  $R^2$  of a model even by adding non-significant or uninteresting terms to it.)

---

<sup>8</sup>Note that the mean and standard deviation for standardized variables won't always be reported by stats software as always equal to exactly 0 and 1, respectively. Sometime there is rounding error or only a subset of the standardized values are being used to re-calculate the mean & standard deviation.

The next table, ANOVA, presents the results of the linear regression in terms of just that. This presents the familiar Sum of Squares for the DBT term (when DBT = 1, as it shows in the left-most column) as well as the Residual (Sum of Squares) which you should now recognize is the unexplained variance. The  $F$ -score and  $p$ -value (Sig.) both indicate the significance of the DBT term.

One more thing to note about the ANOVA table ... is that there even is one: The presence here of an ANOVA table—when we didn't explicitly tell SPSS to run an ANOVA—underlines the fact that what we're doing in a linear regression is the same as we would do in an ANOVA. Again, one reason to prefer a linear regression over an ANOVA is because of the greater flexibility of a linear regression. Of course, if you don't need this greater flexibility, then this also means that running an ANOVA is just fine if that's all you need; in addition, ANOVA source tables may also be more accessible to audiences without quite as much sophisticated understanding of statistics as you now have.

The results of the ANOVA, correlation, and linear regression analyses are all quite similar. Indeed—to the extent that our underlying assumptions hold—the results of all three analyses will become even more similar as the sample size increases. Two things to infer from this that are most relevant here are:

1. **Assumptions matter.** Although some assumptions (monotonicity of the data and that data are independently and identically distributed) tend to be more important than others (strict adherence to normality), knowing how well and in what ways our data meet our basic assumptions affect all analyses we do and all inferences we make from them. This remains true for data of all sizes—even if some assumptions become more important as sample size increases and others tend to become less important (e.g., sampling bias becomes more important; adherence to normality even less).
2. (Nearly all) **linear regressions do the same thing.** The fundamental goal of a correlation, ANOVA, multilevel model, logistic regression, and structural equation model are the same. They all test a linear relationship between the variables by computing a slope, determining that slope is determined by computing a loss-limiting functioning (e.g., ordinary least squares or maximum likelihood), and parceling out variance into that which is explained by the model and that which remains unexplained ("error").

In fact, a main goal of demonstrating the relationship between, e.g., a one-way ANOVA and a zero-order correlation is to show that they can be seen as members of the same family of analyses.

### 8.3.1.3 Semipartial Correlation vs. Multivariate Linear Regression

Remember (e.g., from Section 5.3) that a semipartial correlation removes the effect of a third variable<sup>9</sup> from *one* of the two variables in a correlation. Remember too that as odd as this may seem a thing to do, in fact it's done all the time: It's the basis for having two (or more) predictors in a linear model; the effects of each predictor are (mathematically) isolated from each other so that the effect of one is independent of the effect of the other.

If the two predictors are strongly correlated with each other, then the model will perform quite differently if only one is included versus if both are. Exactly in what way it will "act differently" is hard to anticipate ahead of time, but act differently it will. Let's look at examples of that now.

---

<sup>9</sup>Or the effect of both a third and fourth variable, etc.

### 8.3.2 Conducting a Multivariate Linear Regression Using Forward Term Selection

A *multivariate* linear regression is just a linear regression that has two or more predictors.

1. In SPSS, compute the correlation between ZAll\_EFs\_SR\_Slope and Adult\_Sister\_at\_Home (a dummy variable that indicates whether the teens if they lived with a sister who was over 18 years old). You'll see that  $r_{bp} = -.11$  ( $df = 319$ ,  $p = .048$ ). This correlation is small but significant (and, yes, cherry-picked for this example).
2. Look, too, at the correlations between DBT and both ZAll\_EFs\_SR\_Slope and Adult\_Sister\_at\_Home. The zero-order correlation between DBT and ZAll\_EFs\_SR\_Slope is  $-.165$  and between DBT and Adult\_Sister\_at\_Home is  $.04$ . Of course, this correlation between DBT and Adult\_Sister\_at\_Home is not theoretically interesting since there is no reason to believe that participating in the DBT program really affects how many adult sisters one has or vice versa; nonetheless, it serves well as an example of how linear models change when correlated predictors are both included.
3. Now, let us rerun the linear regression predicting ZAll\_EFs\_SR\_Slope with DBT, but this time also add in Adult\_Sister\_at\_Home. I.e., go to Analyze > Regression > Linear..., put ZAll\_EFs\_SR\_Slope in the Dependent field, and put both DBT and Adult\_Sister\_at\_Home in the Independent(s) field.
4. Under Statistics..., make sure Estimates, Model fit, R squared change, and Collinearity diagnostics are all selected. These play into what we will be doing now.
5. Leave everything under Options the same, viz., leave the Stepping Method Criteria to the default, keep Include constant in equation selected, and Exclude cases pairwise. Most of these inform our investigation here, too.
6. Now, turn your attention to the Method: drop-down menu right under the Independent(s) field. This is the method SPSS will use to add or remove terms to the model. I'll explain this further soon, but for right now, select Forward Selection.
7. Hit OK.

#### 8.3.2.1 Results

##### Variables Entered/Removed and the "Stepwise" Strategy

###### Variables Entered / Removed

Variables Entered/Removed <sup>a,b</sup>			
Model	Variables Entered	Variables Removed	Method
1	DBT	.	Forward (Criteria: Probability-of-F-to-enter <= .050)
2	Adult_Sister_at_Home	.	Forward (Criteria: Probability-of-F-to-enter <= .050)
a. Dependent Variable: ZAll_EFs_SR_Slope			
b. Linear Regression through the Origin			

The Variables Entered/Removed table reports which variables are either entered or removed based on the Method: we selected to determine which variables ought to be selected for our final model.

Let me first explain what is being done here and why before we explore more particularly the methods used.

Second, we could think of the whole model and whether a predictor makes a significant contribution to the overall fit of the model. This latter method will (usually) produce the same results as the former, but has the advantages of both allowing us to test significance in more ways and of allowing us to test and consider contributions more flexibly and precisely.

What SPSS is doing here is based on that second approach. It is trying to build the best model, picking from those we suggested to find the combination that has the largest number of significant terms<sup>10</sup>. Here, the only possible predictors SPSS could choose from are just DBT and Adult\_Sister\_at\_Home. However, when exploring larger sets of data, there may be many more potential variables to add.

The general strategy SPSS presumes we are following is to choose which variable(s) to add or remove from a model based on whether that variable significantly changes the fit of the overall model. One advantage of this strategy is that it considers whether predictors themselves are inter-correlated. If two predictors are strongly correlated, then it's unlikely that both will be added to the model; instead, the one that is more strongly predictive of the criterion will be added and the other one won't make the cut since most of its relationship with the criterion will be accounted for by the other predictor that made the cut first. This tends to create a more parsimonious model that is less affected by multicollinearity and yet is still effective.

SPSS, in fact, offers five methods for deciding which predictors to add or remove<sup>11</sup> given this general strategy:

- Enter: The most theory-driven of the methods, Enter lets one add “block” or set of predictors and then add another block, continuing as one wants. In this way, one could, e.g., first add in all of the variables that are not of direct but that one expects will be important to control for when finally looking at the (other) predictors that are of theoretical interest; in other words, one could create a base model with the first block, and then in the second block start adding predictors of theoretical interest to see if those theoretically-interesting predictors prove to be important after all. This is the method I use nearly exclusively.
- Remove: A similar strategy to Enter, SPSS first starts with all chosen predictors added to the model. It then removes those listed in the first block, then those listed in the second block, and continues until there are no more predictors (except the intercept, if present). Enter is used to test whether adding a block of predicts improves the model; Remove is used to test whether removing a block of predictors worsened the model. It's a subtle distinction, and the choice of which to use is one determines based on one's theory and research questions.
- Stepwise: A method that both adds and removes predictors. SPSS does this in “steps.” In the first step, SPSS starts with no predictors in the model (except the intercept); it then tries out each predictor, seeing which of them would be most significant (has the smallest  $p$ -value<sup>12</sup>); if that predictor with the smallest  $p$ -value is also sufficiently significant<sup>13</sup>, then it is indeed added to the model in that first step.

<sup>10</sup>There are other ways to go about doing this than what SPSS is doing here—ways I prefer—but what we're doing here is easy and still useful strategy to know and use.

<sup>11</sup>More about these methods can be found, e.g., in IBM's [Knowledge Center](#).

<sup>12</sup>Instead of choosing the predictor with the smallest  $p$ -value, SPSS can choose the value with the largest  $F$ -value. A small distinction, but worth a footnote.

<sup>13</sup>The criterion used to decide whether the  $p$ -value is sufficiently significant is determined by the values used under Options... > Stepping Method Criteria. The default is to include a predictor if the  $p$ -value is less than .05 and to exclude a predictor if its  $p$ -value is greater than .10.

In the second step, SPSS first tries out all of the variables still not in the model, chooses the one with the smallest  $p$ -value, and adds that to the model if its  $p$ -value is smaller than the pre-set cut-off. Then, still in the second step, SPSS goes through the predictors that have been added to the model; if any of them have become sufficiently *non*-significant, SPSS removes them from the model. If no predictors were added or removed during the second step (if none met the criteria for entry or removal), then SPSS stops and reports this as the final model.

The third and any subsequent steps follow the same procedure as the second step, explained just above. Note that it is possible (unlikely, but possible) that a variable that was previously removed is later re-entered as SPSS fiddles around finding the best set of predictors.

- Forward: This method is like Stepwise, but in which predictors are only added (if they meet the pre-set cut-off) at each step; none are removed. (This is the method we used here because it's simple and worked for what I wanted to show.)
- Backward: This method is also like Stepwise; here SPSS starts with all chosen predictors in the model, and then only removes any at each step, stopping either when no more meet the pre-set cut-off or when there are no more predictors left in it.

### Model Summary

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.165 <sup>a</sup>	.027	.024	.98788008	.027	8.851	1	317	.003
2	.202 <sup>b</sup>	.041	.035	.98245568	.014	4.510	1	316	.034
a. Predictors: (Constant), DBT									
b. Predictors: (Constant), DBT, Adult_Sister_at_Home									

This is the meat of the output. This table shows the performance of the overall model—not the individual predictors—at each step. The model statistics for the first step are given in the row labeled 1 in the Model column. Notice that the  $R$ -value for the first model (which contains only DBT) is the same as the zero-order correlation between DBT and ZAll\_EFs\_SR\_Slope, again showing the similarity of the processes (again, as long as the same assumptions hold).

Remember that squaring a correlation produces the proportion of variance explained. Similarly the model  $R^2$  (here R Square) indicates the proportion of variance in the criterion explained by this model. The Adjusted R Square is the model  $R^2$  adjusted for the number of terms in the model (predictors as well as intercept—if present—and error); as you can infer from the values here, it reduces the  $R^2$ ; this helps protect against inflated the model  $R^2$  simply adding a bunch of terms that have very little—if any—relationship with the criterion.

Adjusted  $R^2$  also adjusted for the sample size; larger sample sizes are adjusted less since it is argued that they better represent the overall population. Given these adjustments, adjusted  $R^2$

values are best used as descriptive statistics when reporting how much, e.g., your model's results and performance may apply to instances outside of your study—when making recommendations to the field, for example. However, adjusted  $R^2$  does not serve well for comparing between models within your analysis, as we will soon do.

The Std. Error of the Estimate (standard error of the estimate, also called the root mean square error) is the standard deviation of the error term in the model; this simply shows how much residual error (variance) there is in the model. Our  $R^2$  value is small—only accounting for 2.7% of the variance in EF change scores—so it's not surprising that there is a lot of residual error; this column reminds us that that is so.

The R Square Change is how much change there is in the R Square value. For the first model, you'll see that the R Square Change value is the same as the R Square value. Personally, I find this confusing since it's not really a *change* in  $R^2$ , but simply the initial  $R^2$  value. The F Change,  $df$ s, and Sig. F Change for this first row are also simply the significance tests for this first model—not the change in the model. They do show that the model is significant ( $p = .003$ ).

The second row (where Model is 2) indeed shows the change in model fit made by adding Adult\_Sister\_at\_Home to the model that already contains DBT. Now, the R, R Square, Adjusted R Square, and Std. Error of the Estimate values *are* for the whole model (showing that the model with both predictors does account for more of the variance in ZAll\_EFs\_SR\_Slope than the one only containing DBT—even when adjusting for the fact there are simple more predictors there), but the other columns are all analyzing the change in the  $R^2$  value between the first and second models. Although the change in  $R^2$  is small ( $.041 - .027 = .014$ ), it is significant ( $F_{1, 316} = 4.51, p = .034$ ); adding the Adult\_Sister\_at\_Home term improved our model—and thus our understanding of EF changes; DBT and Adult\_Sister\_at\_Home both make significant contributions to our understanding of changes in student-reported executive functioning, and these predictors—although mildly correlated ( $r_{pb} = .04$ )—make otherwise unique contributions to predictions of EF changes.

It's worth reiterating how testing for relationships in this way allows for more precise and nuanced insights into the relationships between our variables. We still test whether both predictors are significant, but do so through how much they contribute to our overall understanding: It's not just if a term is significant, but how much it matters in light of everything else we know.

### ANOVA Table

<b>ANOVA<sup>a</sup></b>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.637	1	8.637	8.851	.003 <sup>b</sup>
	Residual	309.363	317	.976		
	Total	318.000	318			
2	Regression	12.991	2	6.495	6.729	.001 <sup>c</sup>
	Residual	305.009	316	.965		
	Total	318.000	318			
a. Dependent Variable: ZAll_EFs_SR_Slope						
b. Predictors: (Constant), DBT						
c. Predictors: (Constant), DBT, Adult_Sister_at_Home						

Finally something familiar. The ANOVA table presents an ANOVA run on each of the models. This table doesn't, however, show the tests of each of the terms in the model—just the overall model. As such, this table really does little more than what was shown by the  $F$ -scores in the Model Summary table, just above.

## Coefficients

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	.078	.061		1.281	.201	-.042	.199		
	DBT	-.424	.143	-.165	-2.975	.003	-.704	-.144	1.000	1.000
2	(Constant)	-.019	.076		-.249	.804	-.169	.131		
	DBT	-.435	.142	-.169	-3.066	.002	-.714	-.156	.999	1.001
	Adult_Sister_at_Home	.237	.112	.117	2.124	.034	.017	.456	.999	1.001

a. Dependent Variable: ZAll\_EFs\_SR\_Slope

The **Coefficients** table also takes us back to more familiar ground, testing the effects of the predictor terms per se. Given how we've coded our variables, this table is a bit more confusing than it would otherwise be, though: Both DBT and Adult\_Sister\_at\_Home are dichotomous variables, so the Unstandardized Coefficients don't give us the insights that variables that have meaningful units would give; the main insights from the Unstandardized Coefficients is that the intercepts in both models are not significant, meaning that the sixth-grade EF scores for students were not different between those who participated or didn't participate in the DBT program ( $t = 1.28$ ,  $p = .20$ ), even when accounting for whether they had adult sister(s) at home ( $t = -0.25$ ,  $p = .80$ ).

Note, though, that the *beta*-weight for the DBT term in the first model is the zero-order correlation between it and ZAll\_EFs\_SR\_Slope. As the column heading makes clear, *beta*-weights are the standardized regression weights, thus here the semipartial correlation (semipartialing out the intercept). In the second model, the DBT *beta*-weight is  $-.169$ ; this is the semipartial correlation between DBT and ZAll\_EFs\_SR\_Slope; semipartialing out slightly improves—clarifies—the relationship between DBT participation and EF changes (as we knew from the  $R^2$  change tests, above). The zero-order correlation between Adult\_Sister\_at\_Home and ZAll\_EFs\_SR\_Slope is  $-.11$  while the semipartial correlation between them is  $.117$ —stronger and in the opposite direction; the adult sister thing is hard to explain or interpret here, but what we can say is that its relationship with EF changes is certainly mediated by DBT participation, underlining the importance of considering other variables in one's analyses.

The Collinearity Statistics columns report two common tests of (multi)collinearity between predictors in the model:

- Tolerance ranges from 0 to 1, with numbers closer to zero indicating that that variable is stronger related to other predictors in the model. By convention (more than reason), tolerances of less than .1 are seen as problematic and should be addressed, e.g., by removing predictors from the model or explaining why there is such high **multicollinearity**.
- The variance inflation factor (VIF) measure the effect of collinearity on the model. VIFs range from 1 to infinity, and values greater than 10 are typically seen as indicating that collinearity is unignorable affecting the performance of the model. This affect is usually to make the

model terms unstable, meaning we can't speak confidently about not only the absolute values of the  $b$ - or  $\beta$ -weights but that we can't even be sure of significance tests on them.

There is only one predictor in the first model, so the Collinearity Statistics expectedly show there is no collinearity. In the second model, both statistics are remain very good, indicating that the contributions of DBT and Adult\_Sister\_at\_Home are largely independent of each other; the weak correlation between those two variables ( $\phi = .04$ ) makes us expect that any collinearity would be quite small.

### Excluded Variables

Excluded Variables <sup>a</sup>								
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Adult_Sister_at_Home	.117 <sup>b</sup>	2.124	.034	.119	.999	1.001	.999
a. Dependent Variable: ZAll_EEs_SR_Slope								
b. Predictors in the Model: (Constant), DBT								

The Excluded Variables table reports what the statistics would have been for each term if they had indeed been added to it. We only have two predictors and the second predictor was all that was added to the second model, so there is no new information gained here from this table. Has we been conducting a more data-driven investigation into a larger set of variables, this table could be used to look at how the model would have performed under different combinations of predictors, even if SPSS hasn't selected to include them.

### Collinearity Diagnostics

Collinearity Diagnostics <sup>a</sup>						
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	DBT	Adult_Sister_at_Home
1	1	1.430	1.000	.28	.28	
	2	.570	1.585	.72	.72	
2	1	1.937	1.000	.11	.10	.11
	2	.728	1.631	.03	.81	.19
	3	.334	2.407	.86	.09	.71
a. Dependent Variable: ZAll_EFs_SR_Slope						

We had told SPSS to include collinearity diagnostics above, so this table also doesn't present much new, but it does give some more information about what we know already. The Eigenvalue column can be used to ascertain where most of the collinearity in a model resides; this is especially useful if there are more than two variables that share unignorable levels of multicollinearity. In investigating multicollinearity, Eigenvalues greater than 15 are generally seen as important; Eigenvalues less than 1 are always ignorable. The Condition Index essentially measures the cumulative effect of the various sources of multicollinearity; values greater than 15 for condition indices are seen as problematic.



# Chapter 9

## Analyzing Residuals

### 9.1 Set up Your Model

1. Open your dataset in SPSS.
2. Go to Analyze > Regression > Linear.
3. Move your dependent variable into the Dependent box.
4. Move your independent variable(s) into the Independent(s) box.
5. Click Plots and Save (details below), then click OK to run the regression.

#### 9.1.1 Steps to Evaluate Residuals

##### 9.1.1.1 Request Residual Plots

- In the Linear Regression dialogue:
  - Click Plots.
  - In the Y box, enter: ZRESID (standardized residuals).
  - In the X box, enter: ZPRED (standardized predicted values).
  - Check Histogram and Normal probability plot.
  - Click Continue.

##### 9.1.1.2 Save Residuals and Predicted Values

- In the Linear Regression dialogue:
  - Click Save.
  - Under Predicted Values, check:
    - \* Unstandardized
    - \* Standardized
  - Under Residuals, check:
    - \* Unstandardized
    - \* Standardized
    - \* Studentized
  - Optionally under Influence Statistics, check:

- \* Cook's distance
- \* Leverage values
- Click Continue, then OK.

## 9.1.2 Interpretation of Outputs

### 9.1.2.1 Histogram of Residuals

- Evaluates **normality** of residuals.
- Should appear approximately bell-shaped.
- For more rigorous assessment, use the Normal P-P Plot or a Shapiro-Wilk test.

### 9.1.2.2 Q-Q Plot (Normal P-P Plot)

- Points should follow the 45-degree line.
- Deviations indicate non-normality.

### 9.1.2.3 Scatterplot of ZRESID vs. ZPRED

- Assesses **linearity** and **homoscedasticity**.
- Should display a random cloud of points.
  - Patterns (e.g., curves, funnels) suggest violations.

### 9.1.2.4 Studentized Residuals

- Check for outliers, e.g., values above  $\pm 2$  or  $\pm 3$  may be influential.

### 9.1.2.5 Cook's Distance

- Evaluates **influence** of observations on the regression model.
- Values  $> 1$  may indicate problematic cases.

### 9.1.2.6 Leverage Values

- Identifies points with unusual predictor values.
- High leverage + large residual = influential observation.

## 9.1.3 Best Practice Summary

Assumption	Tool in SPSS	Interpretation
<b>Linearity</b>	Scatterplot of ZRESID vs. ZPRED	Random cloud indicates assumption met

Assumption	Tool in SPSS	Interpretation
<b>Normality</b>	Histogram & P-P Plot	Approx. normal shape and diagonal line
<b>Homoscedasticity</b>	Same scatterplot (ZRESID vs. ZPRED)	Equal spread across values
<b>Influence</b>	Cook's Distance, Leverage, Studentized Resid	Values near limits may need closer look

## 9.2 Multivariate Linear Regression with Three Predictors Using Enter Term Selection

Now that we (hopefully!) have some understanding of multivariate linear regression, we can look at a slightly more complex one. Here, we are considering three variables, two of which we know are themselves mildly but significantly interrelated: DBT participation and economic distress ( $\phi = -.08$ ,  $p = 0.48$ ). We'll also consider adult sisters at home to build upon what we did above.

We will also move to what I believe is a more generally-defensible method of building and comparing models. Presuming we are interested in testing if DBT participation affects changes in students' executive functioning, we are probably interested not only if DBT participation matters, but if it matters more than other, theoretically-uninteresting factors—like how many adult sisters one lives with. Controlling for economic distress also removes an important source of variance we couldn't control through experimental design, and so is worth including here for that reason<sup>1</sup>. Therefore, we will add all (both) of the theoretically-uninteresting predictors to the first model. Although we can (and should) investigate the statistics of this first model, it is primarily intended to serve simply as the null model—the baseline of comparison—for testing the effect of DBT participation. Looked at this way, we see if DBT participation predicts changes in EF—while controlling for known sources of variance that could otherwise mislead our interpretation of the effects of the DBT program.

1. To conduct our analyses, again evoke the main linear regression dialogue box, e.g., via Analyze > Regression > Linear...
2. Again let ZAll\_EFs\_SR\_Slope serve as the Dependent(s), but this time first add Economic\_Distress and Adult\_Sister\_at\_Home to the Independent(s) field.
3. Change the Method: to Enter.
4. Click on the Next button just above the Independent(s) field. This allows us to now enter other predictors into what will be the second block of predictors. Add DBT to the now-blank Independent(s) field.
5. Ensure that in the Statistics dialogue box Model fit, R squared change, and Collinearity diagnostics are selected as is Estimates under Regression Coefficients.
6. In the Options dialogue box, makes sure Include constant in equation and Exclude cases pairwise are also selected.
7. Hit OK.

<sup>1</sup>There are really three, general ways to address noise in one's studies: group assignment (e.g., experimental vs. control), randomization, and what we're doing here: adding possibly-confounding variables to a model to isolate their effect on the variables of interest. What we're doing here tends to get short shrift when discussing experimental design, and—in my opinion—that's too bad since we can't always create the groups we want and randomization doesn't always work and isn't always easy to tell if it did.

### 9.2.1 Results

#### 9.2.1.1 Variables Entered / Removed

Variables Entered/Removed <sup>a</sup>			
Model	Variables Entered	Variables Removed	Method
1	Economic_Distress, Adult_Sister_at_Home <sup>b</sup>	.	Enter
2	DBT <sup>b</sup>	.	Enter
a. Dependent Variable: ZAll_EFs_SR_Slope			
b. All requested variables entered.			

The Variables Entered/Removed table summarizes our steps and that we used the Enter method allowing us—not the data—to decide which predictors to add and when.

#### 9.2.1.2 Model Summary

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.196 <sup>a</sup>	.038	.032	.98376623	.038	6.291	2	316	.002
2	.250 <sup>b</sup>	.063	.054	.97281007	.024	8.158	1	315	.005
a. Predictors: (Constant), Economic_Distress, Adult_Sister_at_Home									
b. Predictors: (Constant), Economic_Distress, Adult_Sister_at_Home, DBT									

Given our goals here, this is the most telling part of the output. The base model in step one in fact predict EF changes well (for field-based data). More interesting, though, is that adding DBT to this already-significant model further improves it, increasing the proportion of variance explained by 2.4% (R Square Change = .024), which is a significant contribution here ( $F_{1, 315} = 8.16$ ,  $p = .005$ ).

**9.2.1.3 ANOVA**

<b>ANOVA<sup>a</sup></b>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12.176	2	6.088	6.291	.002 <sup>b</sup>
	Residual	305.824	316	.968		
	Total	318.000	318			
2	Regression	19.897	3	6.632	7.008	.000 <sup>c</sup>
	Residual	298.103	315	.946		
	Total	318.000	318			
a. Dependent Variable: <u>ZAll_EFs_SR_Slope</u>						
b. Predictors: (Constant), Economic_Distress, Adult_Sister_at_Home						
c. Predictors: (Constant), Economic_Distress, Adult_Sister_at_Home, <u>DBT</u>						

The ANOVA table simply reinforces what the Model Summary table contains, showing, e.g., not only that adding DBT made a significant improvement in model fit, but that the model containing all three predictors (Model 2) significantly predicted ZAll\_EFs\_SR\_Slope scores. ( $F_{3, 315} = 7.01, p < .001$ ).

### 9.2.1.4 Coefficients

Coefficients <sup>a</sup>										
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-.294	.099		-2.953	.003	-.489	-.098		
	Adult_Sister_at_Home	.203	.112	.101	1.818	.070	-.017	.424	.996	1.004
	Economic_Distress	.333	.114	.161	2.921	.004	.109	.558	.996	1.004
2	(Constant)	-.207	.103		-2.017	.045	-.410	-.005		
	Adult_Sister_at_Home	.217	.111	.107	1.956	.051	-.001	.434	.994	1.006
	Economic_Distress	.306	.113	.148	2.701	.007	.083	.529	.989	1.011
	DBT	-.403	.141	-.156	-2.856	.005	-.680	-.125	.992	1.009

a. Dependent Variable: ZAll\_EFs\_SR\_Slope

Perhaps the most interesting thing to note from the Coefficients table for this family of analyses is that the DBT term's Beta-weight is lower when including Economic\_Distress (and Adult\_Sister\_at\_Home), reflecting that the shared variance between DBT and Economic\_Distress is also shared by ZAll\_EFs\_SR\_Slope: Some of the effect of the DBT program is mediated by the adolescents' levels of economic distress.

The small levels of collinearity between DBT and Economic\_Distress suggests that the variance they share is nearly entirely itself associated with ZAll\_EFs\_SR\_Slope—that very little of their shared variance is left to create collinearity between them.

### 9.2.1.5 Excluded Variables

Excluded Variables <sup>a</sup>							
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics	
						Tolerance	VIF
1	DBT	-.156 <sup>b</sup>	-2.856	.005	-.159	.992	1.009

a. Dependent Variable: ZAll\_EFs\_SR\_Slope

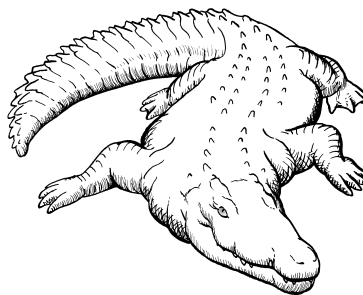
b. Predictors in the Model: (Constant), Economic\_Distress, Adult\_Sister\_at\_Home

Since we chose which predictors to add and when—and since there were only two steps—the Excluded Variables table is again uninteresting.

### 9.2.1.6 Collinearity Diagnostics

Collinearity Diagnostics <sup>a</sup>							
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Adult_Sister_at _Home	Economic_ Distress	DBT
1	1	2.326	1.000	.05	.07	.05	
	2	.480	2.202	.03	.81	.22	
	3	.194	3.462	.92	.12	.72	
2	1	2.556	1.000	.04	.06	.04	.04
	2	.782	1.808	.00	.03	.04	.88
	3	.480	2.308	.03	.81	.22	.00
	4	.183	3.742	.93	.10	.70	.08
a. Dependent Variable: ZAll_EFs_SR_Slope							

The multicollinearity between the predictors is greater in this family of models than it was in the previous family. Nonetheless, it is negligible.





## Chapter 10

# Linear Regression Modeling with SPSS, Part 2: More about ANOVAs and Dummy Coding

### 10.1 Overview

This chapter seeks to further demonstrate how to two correlated predictors can be handled with both an ANOVA and—more generally—with a linear regression model. It also presents more details about conducting an ANOVA and about interpreting dummy variables.

### 10.2 Data

We will use the EF\_Slope\_Data.sav dataset for these additional analyses, focusing on a different set of variables within that data set. We'll now be looking at the effects of both gender and special education status on English / language arts (ELA) grades.

*Please download this data file again from BlackBoard for use here.* These are “[synthetic](#)” data, based on real data but changed to further help ensure the participants' confidentiality. In addition to making it more secure, I have manipulated the data to make the relationships between gender, special education status, and ELA grades stronger for instruction here.

### 10.3 Relationships Between Dummy Variables: Crosstabs and $\chi^2$ Tests

Let us first look at the relationship between gender and special education status.

Both gender and special education status are dummy variables. Gender is set here to indicate whether a student is female, so 0 = not female<sup>1</sup> and 1 = female. Special education status here

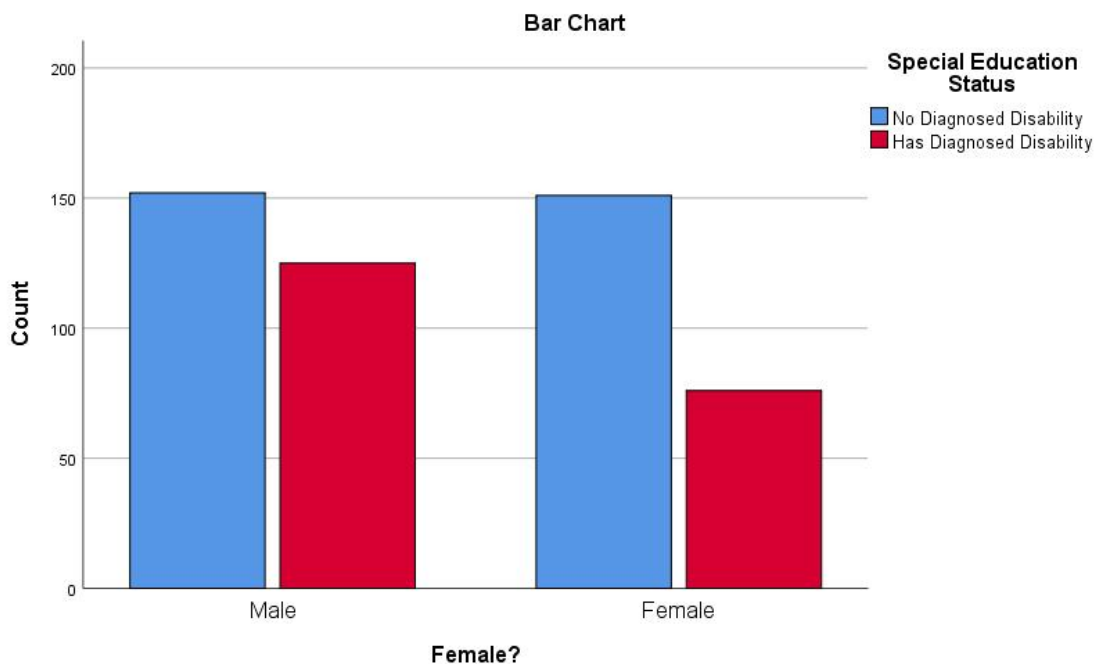
---

<sup>1</sup>In collecting these data, students were asked to indicate whether they were male or female, so gender is dichotomized here.

indicates whether a student has an individualized education program (IEP), so 0 = no IEP and 1 = has an IEP.

We *could* compute a correlation between these two dummy variables (correlations per se are just descriptive), but more information and a more accurate representation of the variable is obtained by looking at a frequency table showing, e.g., how many females have or don't have an IEP.

1. In SPSS, go to Analyze > Descriptive Statistics > Crosstabs
2. Place Gender in the Row(s) field and Spec\_Ed in the Column(s) field. Since they are both nominal, SPSS knows to populate the table with frequency counts.
3. In Statistics... ensure the Chi-square is selected; none of the other options pertain here, so click Continue
4. Under Cells..., in the Counts area, make sure that Observed is selected. The rest of the options in the Cells... dialogue can be interesting, but are pretty straight forward aren't needed here. (Among the other options, I would normally also select Expected under Counts to see for myself how different the actual (observed) counts are from the expected: I can categorically say that it's always a good idea to see the data for yourself and not just rely on a test to tell you what's matters. However, I want to keep the output a bit clean to facilitate interpretation.)
5. Under the list of variables to choose from, select Display clustered bar charts and make sure Suppress tables is *not* selected.
6. Click OK
7. In the Output, let us first look at the Bar Chart. The chart shows exact values, so there is no need for confidence intervals.



Note that we could prettify this chart if we wanted to want to publish it, e.g., by creating better titles, including for the axes.

8. The output starts with a summary of complete and missing data, including the Gender \* Spec\_Ed Crosstabulation ("crosstab")
9. The Chi-Square Tests table contains the following:

1. Pearson Chi-Square<sup>2</sup> is an uncorrected test for whether the counts in the cells differ from "expected" values. "Expected" here means that the proportions are the same, viz., that the proportion of students with IEPs is the same among the boys as it is among the girls. So, there may be more boys than girls, but the *percent* of boys with IEPs is not discernibly different than the percent of girls with IEPs.xxx

Both of these variables have two levels (male/female, has / doesn't have an IEP), so comparisons of counts between them creates a  $2 \times 2$  table. To compute the degrees of freedom for this test, we subtract 1 from the number of levels for each variable. We then multiply those values, or:

$$(2 df_{Gender} - 1) \times (2 df_{IEP Status} - 1) = 1 \times 1 = 1$$

We are therefore testing against a  $\chi^2$  distribution of 1 *df*. The mean and standard deviation of a  $\chi^2$  is determined by the degrees of freedom; more specifically, the mean of a  $\chi^2$  is the *df* and the standard deviation is  $2 \times df$ . So, We are testing this Pearson Chi-Square values against a null  $\chi^2$  with a mean of 1 and a *SD* of 2. Remember that values greater than about 2 *SDs* away from a mean<sup>3</sup> are usually considered significant. Since the mean for this null  $\chi^2$  is 1 and the *SD* is 2, two *SDs* away would be 4 points away from the mean of 1. Therefore, any  $\chi^2$  value that is greater than 5 would be considered significant.

The  $\chi^2$  value here is 7.06, which is indeed larger than the critical value for a  $\chi^2$  for 1 *df*; the counts are significantly different. We could report this by saying, e.g., "The proportion of students with IEPs was significantly different among those who identified as female than among those who did not (Pearson  $\chi^2 = 7.06$ ,  $p > .001$ )" or perhaps more simply: "The proportion of students with IEPs differed between the genders."

Looking at the bar chart we generated shows us more clearly what this difference is: Fewer girls have been diagnosed with disabilities warranting IEPs than boys, and we could certainly report that instead. The Pearson  $\chi^2$  (and other  $\chi^2$  tests here) are inherently non-directional<sup>4</sup>, but we can use that figure (or the table) to argue what this difference is.

Having now seen how the proportion of IEPs differed, we could instead describe this and the  $\chi^2$  test as, e.g., "A larger proportion of males students had IEPs than did female students (Pearson  $\chi^2 = 7.06$ ,  $p > .001$ )."

2. Continuity Correction reports the Yates' correction. This is only presented in SPSS  $2 \times 2$  tables (and is only appropriate for such tables), like we have here. The Pearson  $\chi^2$  test tends to be biased "upwards" meaning it is overly optimistic and generates Type 1 (false positive) errors. Yates' correction attempts to adjust for this by making the test more conservative. You'll see here that the Value (the  $\chi^2$ ) under Continuity Correction is slightly smaller than that under Pearson Chi-Square. Both of these tests are appropriate when all cell counts are greater than 10 (some suggest greater than 5), and the Yates' correction is generally advised.

<sup>2</sup>Karl Pearson is the person who first devised using  $\chi^2$  distributions in statistics, so SPSS calls this a *Pearson  $\chi^2$* . This is nonetheless just the same  $\chi^2$  we used anywhere else. In other words, it's redundant—or superfluous—to call this a "Pearson  $\chi^2$ " instead of just " $\chi^2$ ".

<sup>3</sup>Assuming it's a normal or  $\chi^2$  distribution—or one of the other distributions that are like them, such as the *t* or *F* distributions used to test *t*- and *F*-scores.

<sup>4</sup>In fact, it's a one-tailed test, testing whether the proportion (viz., of IEPs) is the same or different, but it doesn't test whether any differences in the proportions are due to larger or smaller proportions here. For our uses—and likely any you will encounter—considering it a non-directional test of differences *somewhere* suffices.

3. The Linear-by-Linear test is the Mantel-Haenszel test, which is useful when one wants to look at the association between two nominal variables while controlling for the effect of a third variable—akin to partialing out that third variable. We're not doing that here, though, so this statistic isn't interesting.
4. The Likelihood Ratio test, also called the G-test, uses odds ratios to determine the likelihood that the given frequencies in the cells occur by chance. As you can see, this computes a very similar value to the unadjusted (Pearson)  $\chi^2$  value. We won't consider this any more here, but will revisit likelihood ratios when we look at tests of whole linear models.
5. Fisher's Exact Test does not have a statistic like a  $\chi^2$  that is computed; it is simply a probability test of those frequencies themselves. Here as with the  $\chi^2$  tests, a  $p < .05$  (or whatever one sets for significance) indicates a significant difference in the actual cell counts from the expected.

We have seen that whether a student has an IEP depends in part on the student's gender. In other words, gender and IEP status are related; they share variance. Therefore, when I talk, e.g., about IEPs, I know I should also consider gender.

### 10.3.1 Relationships with ELA Grades

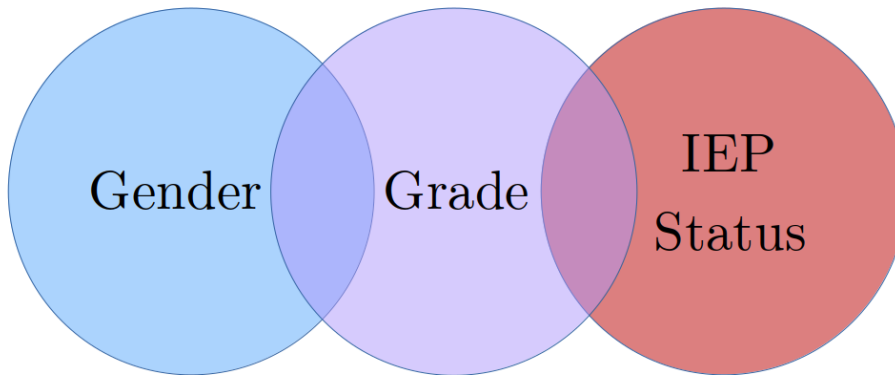
Let us now investigate whether gender and IEP status are related to students' grades.

1. Go to Analyze > Correlate > Bivariate... and add Gender and ELA\_Grade to the Variables field. We'll be computing point biserial correlations (nominal vs. continuous) which are computationally equivalent to Pearson's correlations, so leave that option selected under Correlation Coefficients.

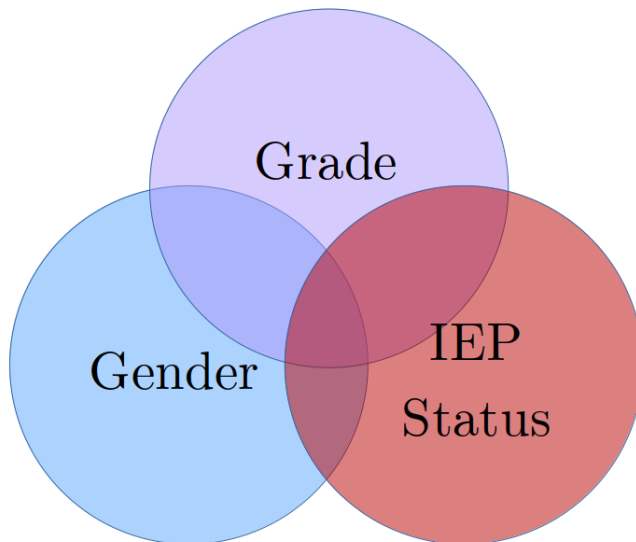
The resultant output shows that  $r_{pb} = .103$  ( $n = 204$ ,  $p = .144$ ). These variables are not significantly correlated here, and gender accounts for ~1% ( $.103^2 = .0106 \approx .01$ ) of the variance in ELA grades, what Cohen (1988) would call a "small" effect (q.v., Chapter 6).

2. Looking now at IEP status, let's remove Gender from the Variables field in Analyze > Correlate > Bivariate..., leave in ELA\_Grade, and add Spec\_Ed.
3. The correlation is even stronger,  $r_{pb} = -.36$ . IEP status accounts for about 10% ( $-.36^2 = .13$ ) of the variance in one's grade. However, from our crosstabs work above, we know that the variance in IEP status is itself related to a student's gender. In other words, some of that .13 variance is due to gender.

So, we know that some of the variance in IEP status is due to gender. We also know that some of the variance in ELA grades is also due to gender. We don't yet know, however, if the effect of gender on IEP status is from the same aspects of gender as the effect of gender on grades. The shared variances between these three variables could kind of be like this:



Or perhaps more like this<sup>5</sup>:



A little less abstractly, one reason why boys tend to be diagnosed with disabilities more often than girls is because boys tend to “act out” more than girls: Boys display externalizing behaviors more frequently and intensely than girls, and this encourages schools to try to figure out ways of helping the boys be less disruptive. Girls tend to suffer in silence.

However, it may well be that acting out isn’t what it is about being a boy that affect his grades. A boy may be the class clown or trouble maker, but he may be quiet bright and do well despite his disruptions of others—or at least attracting attention to oneself may not be what gets one a particular grade. Indeed, boys of any level of disruptiveness tend to be praised for successes in math courses while girls tend to be praised for successes in ELA courses—even the out-spoken ones.

We will next look at the relationships between these three variables. We’ll first look at them through an ANOVA; the ANOVA should help set the stage since this is an analysis you’ve become familiar with and since this is a very common analysis used.

After we review the ANOVA, we’ll look at the relationship through a more general linear regression analysis. We’ll see how it’s similar to an ANOVA and how it differs. The overall goal here is to

---

<sup>5</sup>A moment’s reflection will reveal that my Venn diagrams aren’t really reflective of the point I’m trying to make, but I decided to go with a simplified representation that hopefully still works.

help you learn the pros and cons of analyses you're familiar with (viz., ANOVAs & *t*-tests) and the reasons consider times to use some other linear regression analyses.

## 10.4 Using an ANOVA to Predict ELA Grades with Gender & IEP Status

### 10.4.1 ANOVA Review

Remember that an ANOVA is used to test whether one or more nominal variables (IVs) significantly predict a continuous outcome variable (DV). More specifically, an ANOVA tests whether the mean outcome score differs between one or more of the levels of a predictor. (Here, for example, if the mean ELA grades differ between girls and boys.)

The ANOVA itself can't say *which* levels of the predictor are different, though. (For example, it can say that there is a significant effect for gender, but not which has greater scores). To find which levels differ, we usually conduct a post hoc analysis.

Done well, a reason conduct an ANOVA first is to help control for Type 1 errors: Instead of running a whole bunch of pairwise tests between all levels of a variable (in all variables added to the ANOVA)<sup>6</sup>, we first run a few, overall tests. We then only conduct post hoc tests on variables that the ANOVA found significant, further limiting the number of tests we run and thus the chances of a false positive effect<sup>7</sup>.

Of course, since we only have two levels for each of the predictors here, the ANOVA can tell us if there is a significant difference and then we can simply look at the variable's means to see which is greater.

### 10.4.2 Graphical Review of the Variables

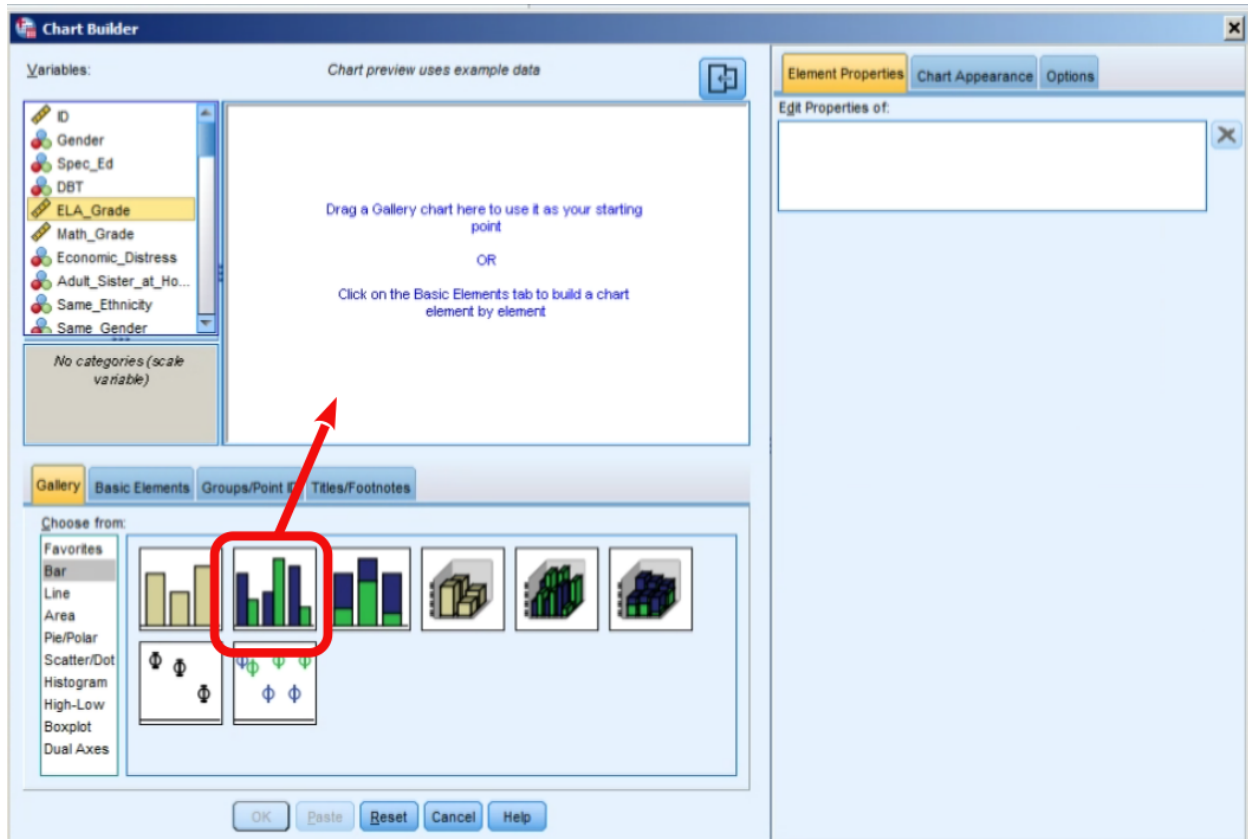
Let's start with looking at graphical representations of these variables and then explore them through an ANOVA.

1. SPSS's Graph > Chart Builder interface is quite useful, even if spreadsheet programs like Excel & Calc have mostly caught up to it.
2. In that dialogue box that opens, drag the beige bar graph near the bottom under Choose from into the main window under Variables:

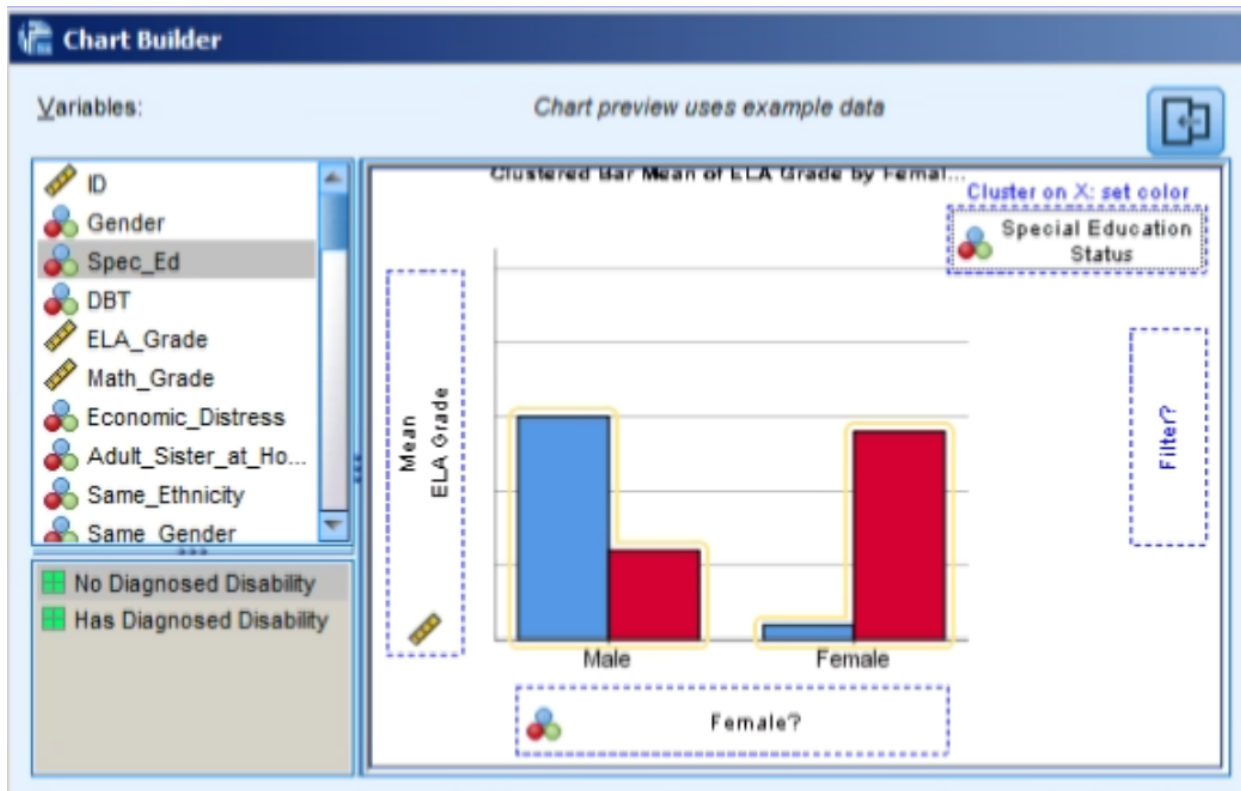
---

<sup>6</sup>And yes, reducing Type 1 errors by conducting fewer significance tests can be seen as an ironic reason to conduct an ANOVA since it's pretty common for researchers to run post hoc analyses on many or all of the nominal variables, and thus end up conducting *more* analyses after counting un-necessary post hocs.

<sup>7</sup>Kao & Green (2008) provide an excellent review both of ANOVAs in general and of the uses of the several post hoc analyses.



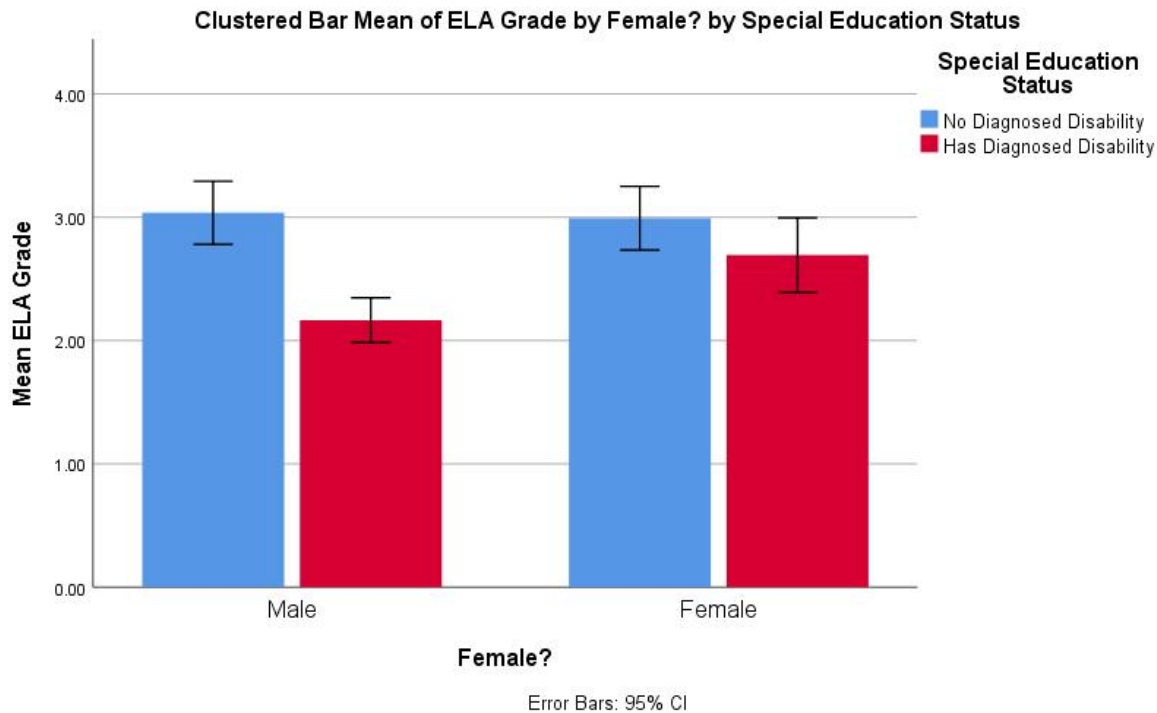
3. Drag ELA\_Grade to the Y-Axis? box in the bar graph that appears in the main window, drag Gender to the X-Axis?, and drag Spec\_Ed to the Cluster on X: Set colors area to the upper right of the graph. That main area should now look like this:



4. The current bar graph purposely resembles the one we created previously while building our crosstabs. However, then we presented absolute counts whereas now we're showing means, so it's worth also showing how well the means represent the sample. In the Edit Properties of: area, select Bar1. The area below that will change, allowing you now to select to Display error bars. Confidence intervals is selected by default; leave that selected, and leave Level (%) set to 95, the value every social scientist (and—more importantly—reviewer) knows and loves.
5. Both the Element Properties and the Chart Appearance tabs have reasonable sets of options for customizing figures, but more can be done when it is generated in the Output window and via syntax.
6. Ways of handling missing data appear under the Options tab. Excluding User-Missing Values is nearly always advisable—the only time I can think to Include them is if you want to report information about the missing cases in the figure.

Under Summary Statistics and Case Values, select to Exclude variable-by-variable, which is tantamount to excluding missing data pairwise instead of listwise.

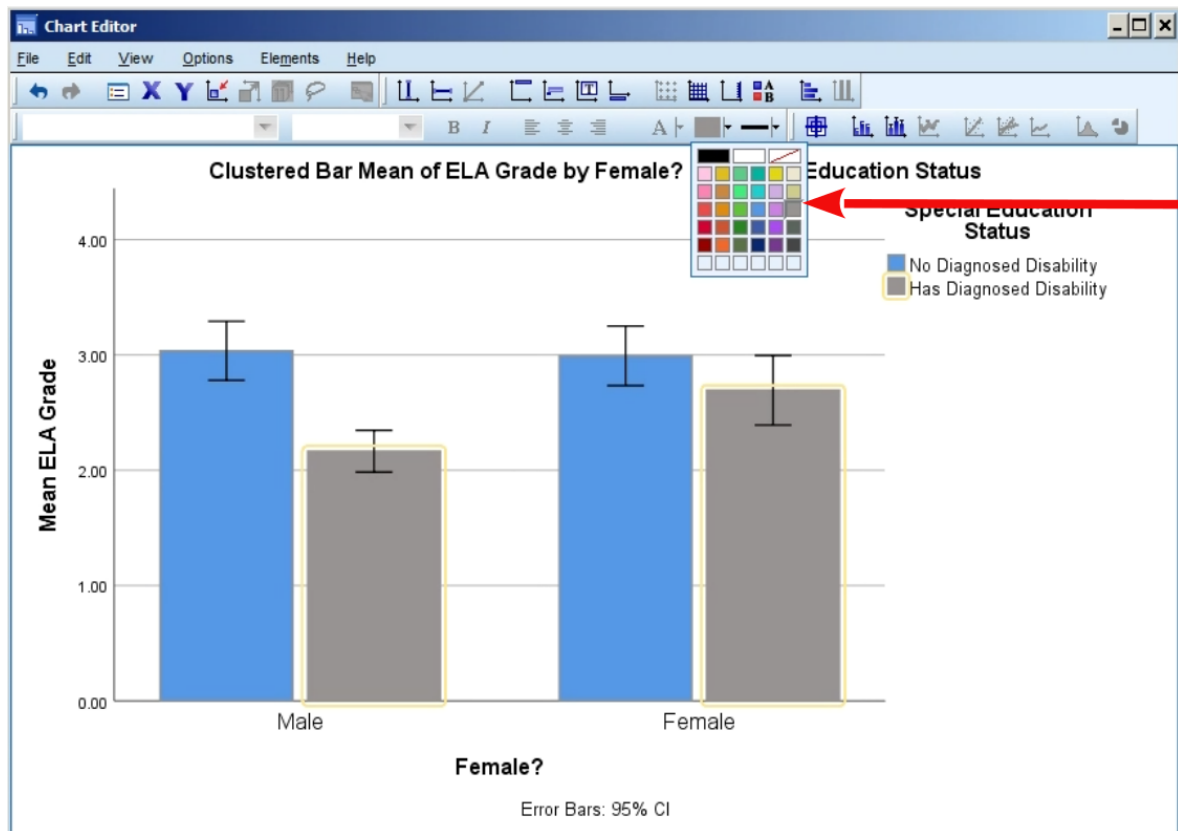
7. Clicking OK will generate this figure:



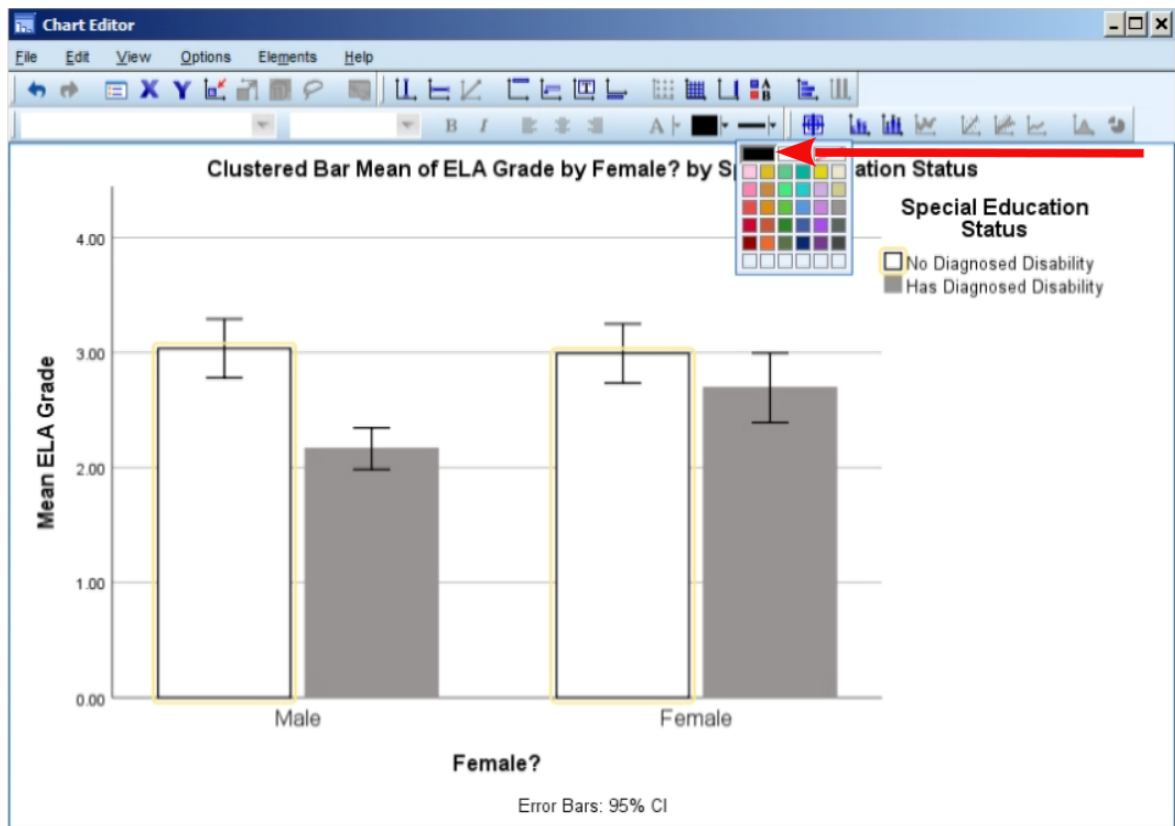
8. The figure shows the difference IEP status made and that girls may have had higher ELA grades than boys—at least among those without IEPs. The 95% confidence interval bars suggest which of these differences are significant<sup>8</sup>.
9. In the Output window, we can modify the figure more. Double-click on it in the Output, to open the figure up in another window with many (Excel-like) options to modify parts of it.
10. The bar chart will now appear as well in its own window; in that window, single-click on one of the No Diagnosed Disability bars (double-clicking can highlight all of the bars, including the Has Diagnosed Disability ones).
11. In the menu bar, choose to change the Fill Color to, e.g., grey<sup>9</sup>:

<sup>8</sup>It's worth noting as well that there is a growing trend—at least among leading statisticians if not general researchers—to rely more on more on less definitive measures like confidence intervals to convey one's results than to rely on up-or-down significance tests. Right now, at least though, my experience has been that reviewers are not comfortable when one excludes significance tests, so I recommend presenting data both with, e.g., confidence intervals and with *p*-values. Note, too, that confidence intervals and *p*-values are not equivalent: The confidence intervals are computed making few assumptions about the data and do not consider, e.g., if the data are skewed unless you modify the intervals in light of skewness. If you want to account for skewness, the preferred method is currently to use bootstrapped intervals as describe in, e.g., Visalakshi & Jeyaseelan (2014) or to use [log transformations](#). The latter is more well-known and accepted, but the former is likely preferred since it doesn't make any assumptions about the underlying population and is more generally use-able whereas log transformations are only useful for data that are nearly normal but simply skewed.

<sup>9</sup>Colors can be useful—and sometimes necessary—but grey scales print in hard copies well and often are more easily seen by people with colorblindness.



We could also change the No Diagnosed Disability bars to white and increase the thickness of the borders (with the next menu item to the right) to create a slightly more manuscript-ready figure:



We can similarly change the fonts, the title contents, etc. Note that once you've tweaked your figure to your (and your committee's) liking, you can click on File > Save Chart Template to create a template that you can later File > Apply Chart Template to other figures to create a nice, consistent look.

### 10.4.3 Using an ANOVA to Test Variables

#### 10.4.3.1 Generating the ANOVA Model

1. SPSS categorizes ANOVAs under general linear models (Analyze > General Linear Models)<sup>10</sup>, which can be taken to emphasize that they are a type of linear regression. The *Univariate* option under Analyze > General Linear Models is for any model that has one outcome (criterion) variable; *Multivariate* is for when there are more than one criterion (e.g., a MANOVA). We have one criterion, ELA\_Grade, so choose Univariate and add ELA\_Grade to the Dependent Variable field.
2. Choosing whether to place predictors in the Fixed Factor(s) field or the Covariates field matters affects the assumptions that are made by the model about that variable and a bit how we interpret the results. It suffices to say that nominal variables should be added as fixed factors and that ordinal, interval, and ratio variables should be added as covariates<sup>11</sup>. Since both of our variables are fixed factors, place Gender and Spec\_Ed in the Fixed Factor(s) field.

<sup>10</sup>The naming of analyses gets capricious and confusing from there, though. A "general linear model" is a type of "generalized linear model." Multilevel models and logistic regression are also types of *generalized* linear models, but let's leave it at that. The list of terms in Table C.1 in Appendix B is intended to help clarify this and other confusions.

<sup>11</sup>The non-simplified answer is that designating a variable as a fixed factor means we're assuming that all possible levels of that variable are present. Both our variables are fixed factors since we have dichotomized them into "Is female" or "Is

3. By default, SPSS adds in interaction terms for all fixed effects. (So, if we had three fixed factors—say A, B, and C—SPSS would include the  $A \times B$ ,  $B \times C$ ,  $A \times C$ , and  $A \times B \times C$  interactions.) We do indeed want to look at both the main effects and the gender  $\times$  IEP status interaction, so we want a “full” model. Even though SPSS would create that by default, let’s build it anyway just so you can see how to do it (and thus how to build other models):

1. Under the Model dialogue, first click on the Build Terms button.
2. Change the Build Term(s) Type to Main Effects and then move both Gender and Spec\_Ed to the Model field.
3. Now change the Build Term(s) Type to Interactions. With *both* Gender and Spec\_Ed selected, click on the arrow under Type to create a Gender  $\times$  Spec\_Ed interaction term. That dialogue box should now look like this:

---

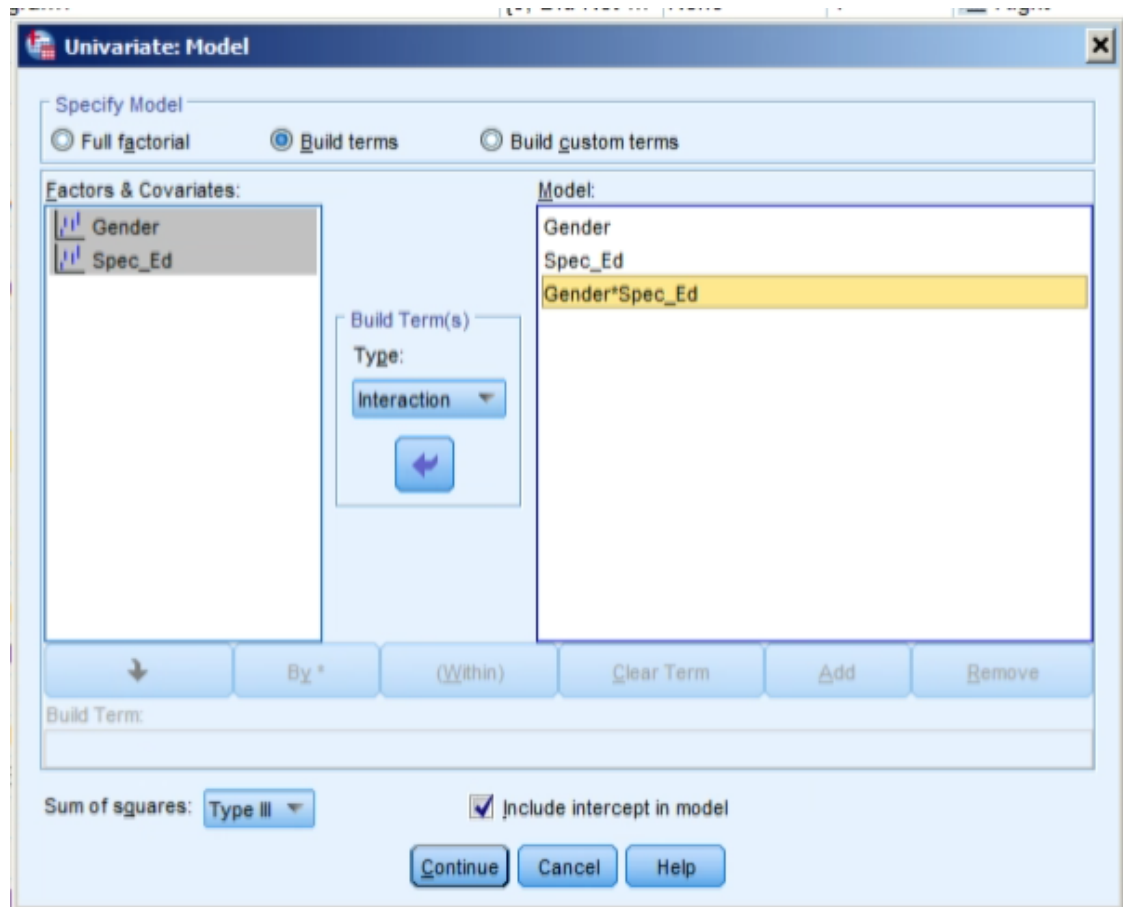
not female” (and “Has an IEP” or “Doesn’t have an IEP”). Whenever the levels of our variable exhaust all possible options, then a variable is fixed. The null we’re testing against is that the means are the same for all levels of the variable.

With *random factors*, we’re assuming that not all levels are present in our data. For example, we may be testing differences between hospitals: We have data from a few hospitals, but certainly not *all* hospitals. The null hypothesis we’re testing against is that there is no variance between any levels in the population (i.e., it’s a test against inferred population variance, not differences in the sample means).

For covariates, we are computing the slope of that variable with the criterion—whether the slope differs from zero. Since we compute a slope, we could estimate the values on the criterion for values of the covariate that were not included in the model.

It is bit unfortunately that SPSS calls this a covariate since we often think of a covariate as something we are controlling for in a model—something we want to partial out so that we can see the effect of another variable more clearly. Covariates here can certainly be used to do that, but they don’t need to be: Variables placed in the covariates field can be interpreted as the main variables of interest and the fixed ones could be ones we’re partialing out. The math is the same regardless of which term we’re interpreting as the variable of interest and which we’re adding to the model to partial out its effect.

A further point to make is that SPSS doesn’t compute random factors efficiently in Analyze > General Linear Models. It would be better to use the Analyze > Mixed Models > Linear dialogue for models with variables that are continuous. Nonetheless, this isn’t absolutely necessary to do, and the output you get from adding random factors here won’t likely ever greatly differ from the results gained from the Mixed Models analyses.



4. Let me reiterate that, by default, SPSS creates a full factorial model for all fixed factors, so we didn't need to do this here (and could have done it more automatically through this dialogue). I did want to show you how so you can modify your models term-by-term rather easily through this particular dialogue.
4. The Contrasts dialogue lets us determine if and how SPSS tests differences between levels of the variables. The default is to compare None (and since ours are dichotomous (dummy) variables, any effect of a variable is a difference between those two levels). The options are explained in more detail, e.g., [here](#), but suffice it to say that Deviation—in which each level is compared against the overall mean—is common, that the other options depend which differences matter most of a given study, and that contrasting differences between levels is often better handled via post hoc analyses anyway.
5. Plots would allow us to create figures quite like we did with Graphs > Chart Builder, but with fewer options made though a more streamlined interface.
6. The Post Hoc dialogue allows one to compute those. Again, Kao & Green (2008) [provide](#) nice, terse explanations and recommendations of commonly-used ones.
7. The EM Means dialogue is useful for our purposes here. This area lets us generate estimated marginal means; these are the means for one factor when other variable(s) are partialled out.
8. Under Options, please select Estimates of effect size and Observed power.
9. Click OK.

### 10.4.3.2 ANOVA Output

1. After reporting the numbers of cases for each variable, SPSS outputs the source table as Tests of Between-Subjects Effects:

**Tests of Between-Subjects Effects**

Dependent Variable:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power <sup>b</sup>
Corrected Model	21.690 <sup>a</sup>	3	7.230	12.493	0.000	0.195	37.479	1.000
Intercept	1146.481	1	1146.481	1981.028	0.000	0.927	1981.028	1.000
Gender	2.276	1	2.276	3.934	0.049	0.025	3.934	0.504
Spec_Ed	13.229	1	13.229	22.858	0.000	0.129	22.858	0.997
Gender * Spec_Ed	3.169	1	3.169	5.476	0.021	0.034	5.476	0.643
Error	89.703	155	0.579					
Total	1265.810	159						
Corrected Total	111.393	158						

a. R Squared = .195 (Adjusted R Squared = .179)

b. Computed using alpha = .05

2. A familiar sight (I hope), we can see from this table that all of the terms—the intercept, gender, IEP status, and the gender  $\times$  IEP status interaction—are all significant at  $\alpha = .05$ . Now, however, a few other parts of this table are of interest (and others perhaps simply worth explaining / refreshing):

1. The Corrected Model term is a test of the whole model—yes, like we are doing with linear regressions.
2. The Type III of Sum of Squares indicates how the terms were added to the model. The math is a bit eldritch (even I have to look it up to remember it), but a summary should suffice. The type used here, Type III, is computed by having all of the terms added to the model at once so that their variances are computed in light of all other terms; in essence all terms are partial regressions, even the intercept and interaction. Given our interests here in partial regressions, this is appropriate<sup>12</sup>.
3. The Partial Eta Squareds are measures of effect sizes for the given terms. As researchers move uneasily away from up-or-down significance tests, they are often using effect sizes as rather sturdy canes for support. Personally, I'm among them, and I report them even as I regularly report  $p$ -values, e.g., as “The main effect for gender was significant ( $F_{1,155} = 2.28$ ,  $p = .049$ ,  $\eta^2 = 0.025$ ).”  
A bit ironically, people have looked for “tests” of effect sizes. Nearly always, this is a hearkening to the original work on them by Jacob Cohen (1988), where he [suggested](#) for  $\eta^2$  that .1 could be considered “small,” .25 could be considered “medium,” and .4 considered “large”<sup>13</sup>. By that standard, gender has a rather small effect that is nonetheless significant here.

<sup>12</sup>Type II sum of squares is similar in that the terms are all added together, but the main effects are partialled in light of each other but not in light of the interaction; if there are no interaction terms, then Types II and III are computational the same. In Type I, the terms are each added one after the other, like we did in the last handout for the linear regression model; the order they are entered is the order they're listed in the Fixed Factor(s) field, and then the Random Factor(s) field, and finally the Covariates(s) field.

<sup>13</sup>Please see Chapter 6 for more on effect size and guidelines for “small,” “medium,” and “large.”

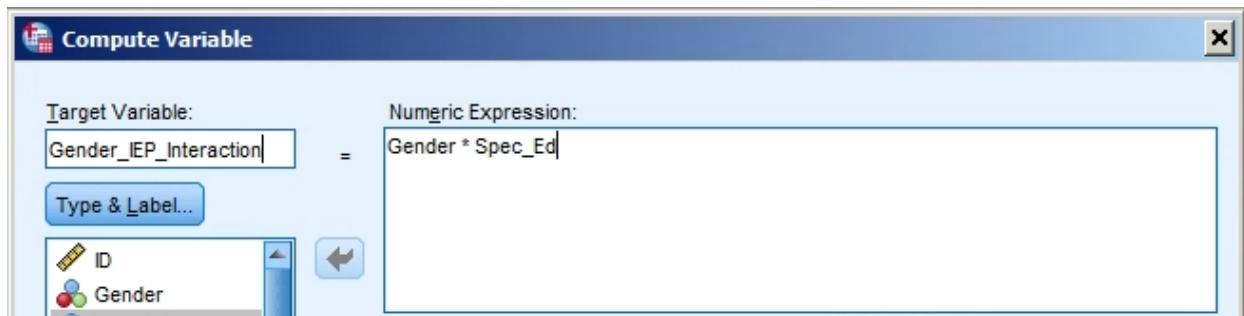
4. The **Observed Power** is the **estimated power** of the  $F$ -test of the given parameter based on the data. It is the estimated probability that a real effect would be detected; it is affected by the sample distribution and size. SPSS computes the Observed Power with the Noncent. Parameter, a statistics that follows a, well, non-central (skewed) distribution that is a composite of  $\chi^2$  and Poisson distributions. In other words, don't worry about it, just know it's used to compute Observed Power—which itself is really a **useless statistic**.
  5. Most importantly right now, note that the  $R^2$  reported under the table is .195 and that the adjusted  $R^2 = .179$ <sup>14</sup>.
3. Although we could generate parameter estimates and marginal means (means for variable levels adjusted for other variables in the model), they are easier to interpret when we compute the results through a linear regression which we will do now.

## 10.5 Linear Regression

### 10.5.1 Creating an Interaction Term

We will use the same model for a linear regression that we did for an ANOVA. SPSS doesn't automatically compute interaction terms for linear regression models like it does ANOVAs. Fortunately, this is quite easy to do:

1. Click on Transform > Compute Variable.
2. In the Target Variable field, type, e.g., Gender\_IEP\_Interaction.
3. Move Gender to the Numeric Expression field.
4. Click on the \* (asterisk) button in the “number pad” below the Numeric Expression field.
5. Move Spec\_Ed to the Numeric Expression field. The top of that dialogue box should now look like this:



6. Click OK to create this variable. It will appear at the far end (right of the Data View, bottom of the Variable View) of the data matrix; you may want to move it to the left / top of the set for easier access.

Yes, all we did was multiply Gender by Spec\_Ed. That is all an interaction term is: the two variables multiplied by each other<sup>8</sup>. For a dummy variable like this, of course,  $0 \times 0 = 0$ ,  $1 \times 0 = 0$ ,  $0 \times 1 = 0$ , and  $1 \times 1 = 1$ , so the values for this interaction term are all zeros except when for females (Gender = 1) who also have IEPs (Spec\_Ed = 1). I'll explain this later, but simply note it now.

<sup>14</sup>Remember that adjusted  $R^2$  is adjusted for the number of terms in the model since having more terms—even non-significant ones—can increase the model  $R^2$ .

## 10.5.2 Computing a Linear Regression with an Interaction Term

### 10.5.2.1 Generating the Linear Regression Model

To present a similar model to the ANOVA above, let's enter all of the terms together.

1. Click on Analyze > Regression > Linear.
2. Enter ELA\_Grade in the Dependent field and Gender, Spec\_Ed, and Gender\_IEP\_Interaction to the Independent(s) field (and setting the Method is Enter).
3. Under the Statistics area, make sure Model fit and R squared change are both selected.
4. Under the Options... area, make sure Include constant in equation and Exclude cases pairwise are selected.
5. Click OK.

### 10.5.2.2 Linear Regression Output

1. The Model Summary shows that the model does account for significant amount of the variance in the data ( $F_{3, 157} = 12.6$ ,  $p > .001$ ):

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	.441 <sup>a</sup>	0.195	0.179	0.67781	0.195	12.643	3	157	0.000

a. Predictors: (Constant), Gender\_IEP\_Interaction, Gender, Spec\_Ed

Also note the the  $R^2$  and adjusted  $R^2$  are the same as we found with the ANOVA: This is the same model, just looked at in terms of the model fit instead of the significance of model parameters.

2. The ANOVA table in the linear regression about shows similar statistics to the Corrected Model row in the source table for the ANOVA above: In the source table above, the  $F$ -score for the Corrected Model was 21.69; here, the similar statistic is the  $F = 12.643$  in the Regression row:

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17.426	3	5.809	12.643	.000 <sup>b</sup>
	Residual	72.130	157	0.459		
	Total	89.556	160			

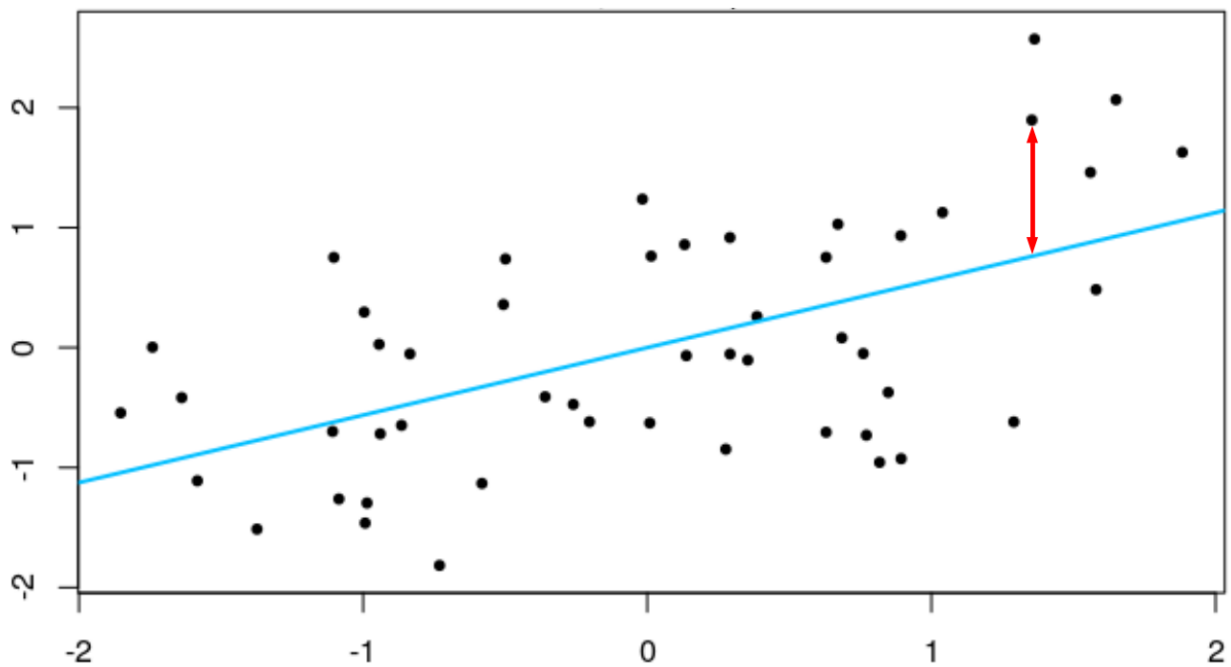
a. Dependent Variable: ELA\_Grade

b. Predictors: (Constant), Gender\_IEP\_Interaction, Gender, Spec\_Ed

This is because the sums of squares are computed a bit differently here. Nonetheless, the outcome is the same.

To show that the outcome is the same, remember that the model  $R^2$  is the proportion of total variance in the data that is accounted for by the model; in other words,  $R^2$  is the variance in the model divided by the total variance.

Now, remember what a “sum of squares” here is: It’s the squared differences between the expected value and the actual value, all added up. So, if the blue line in the figure below is the regression line estimated by an entire model (not this ELA model, but just a made-up one between two z-score variables):



Now, if we didn’t have that regression line to help us—if we had no information except the column of ELA grades—then the best guess we could make about the grade for each student would be the mean ELA grade for the whole sample. In that figure, both variables are z-scores, so the means are zero: If I didn’t use the values of the predictor to estimate that line, then the best guess we would have for that person’s score on the criterion would be the mean, zero. In this case, to get the “sum of squares,” we’d first get the difference of each predictor from the mean, then square and sum those values—this would be the sum of squares if we didn’t use any information in the predictor(s): This would be the Total Sum of Squares in the table: 89.556.

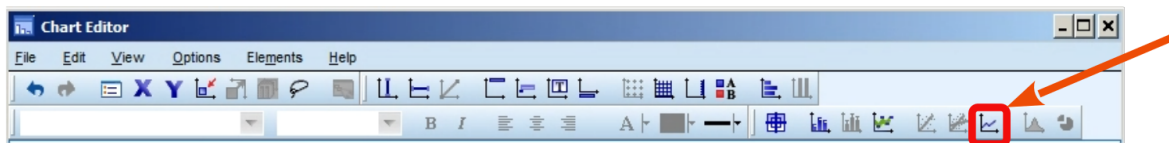
Then the red line shows one of the distances from an actual data point from that estimated line. If we squared this distance—and all of the distances of the dots from the line—and then added up those values, we would get the Regression (or, computed slightly differently, the Corrected Model) Sum of Squares, which here is 12.643.

Remember that the  $R^2$  is the model sum of squares divided by the total sum of squares: Here, that is  $17.426 / 89.556 = 0.195$ . In the ANOVA we computed above, this is  $21.69 / 89.556 = 0.195$ . They both produce the same  $R^2$  value.

### 10.5.2.3 Scatterplot of the Data

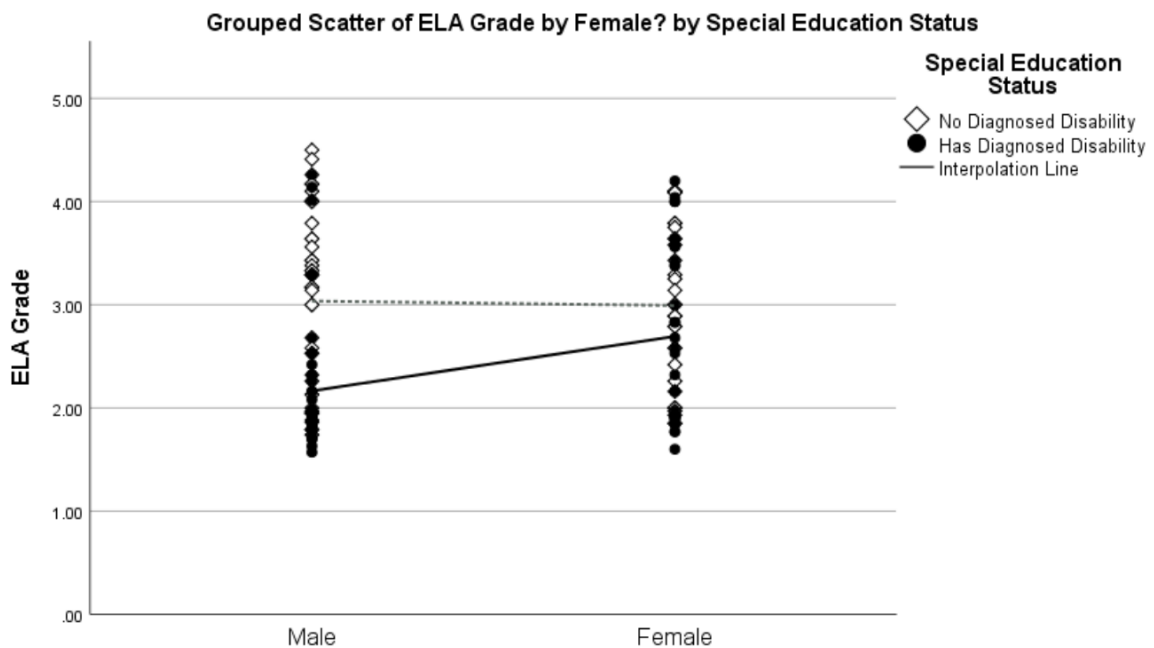
What's that? You say you'd rather see what a chart for these data would look like than an mock one of two made-up z-scores? Well, all right then:

1. Click on Graphs > Chart Builder and select Scatter/Dot from the Gallery tab in the bottom left corner. (You may well see a warning dialogue when opening the Chart Builder saying that "Before you use this dialog, measurement level should be set properly..."; this is a good thing to check, and has been for these data, so it's fine to click OK.)
2. Drag the first figure image, Scatter Plot, up into the main field just under where it says Chart preview uses example data.
3. Just as we did for the bar graph at the beginning of this handout, put ELA\_Grade in the Y-Axis field, Gender in the X-Axis. Also add Spec\_Ed into the Set color? field in the top right.
4. Click OK. The default difference in color
5. Double-click on the figure that's generated, and click on the Add interpolation line button, which is the third button from the right:



This will add a regression line for the IEP status of the males and another for the IEP status of the females. It's not so easy to tell, but the female's line is the upper one.

6. After clicking on elements and using either the tool bar or Properties dialogue (accessed, e.g., via  $\text{Ctrl} + \text{T}$ ), we can create a figure like this:



showing that having an IEP has more of an effect on males' than females' ELA grades (indeed, there is little overlap between the grades of males with and without IEPs, unlike the females).

It also shows that there is a wider range of grades among the males—including that the best (and worst) ELA grades were earned by males.

7. The output also provides the coefficients for the model terms:

### Coefficients<sup>a</sup>

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.983	0.101		29.528	0.000
	Gender	-0.212	0.140	-0.142	-1.516	0.132
	Spec_Ed	-0.870	0.146	-0.571	-5.950	0.000
	Gender_IEP_Interaction	0.804	0.236	0.369	3.399	0.001

a. Dependent Variable: ELA\_Grade

#### 10.5.2.4 Use of Dummy Variables to Estimate Outcomes

The coefficients for the model terms allow us to estimate the group means—and (hopefully) help explain a bit more about dummy variables. The equation for the linear model we analyzed can be written as:

$$\text{Estimated ELA Grade} = \text{Intercept} + \text{Gender} + \text{IEP Status} + (\text{Gender} \times \text{IEP Status})$$

or a bit more abstractly as:

$$\text{ELA\_Grade}' = b_0 + b_1\text{Gender} + b_2\text{Spec\_Ed} + b_3\text{Gender\_IEP\_Interaction}$$

where the tiny apostrophe (') at next to ELA\_Grade denotes that we are estimating—predicting—ELA\_Grade, not simply reproducing it. So, the better the model, the more the predicted scores will replicate the actual ones.

In that second equation,  $b_0$  is what is given in the Unstandardized B column<sup>15</sup> of the (Constant) row of the Coefficients table, so we could rewrite that equation as:

$$\text{ELA\_Grade}' = 2.983 + b_1\text{Gender} + b_2\text{Spec\_Ed} + b_3\text{Gender\_IEP\_Interaction}$$

We can similarly fill in the values for  $b_1$ ,  $b_2$ , and  $b_3$  from the Unstandardized B column to produce:

$$\text{ELA\_Grade}' = 2.983 - 0.212(\text{Gender}) - 0.870(\text{Spec\_Ed}) + 0.804(\text{Gender\_IEP\_Interaction}).$$

Now, remember that Gender, Spec\_Ed, and Gender\_IEP\_Interaction all have values of only either 0 or 1. Gender\_IEP\_Interaction is 1 if the student is a female (Gender = 1) with an IEP (Spec\_Ed = 1); otherwise it's a 0 since  $0 \times 1 = 0$ ,  $1 \times 0 = 0$ , and  $0 \times 0 = 0$ .

So, if we want to estimate the ELA\_Grade score for a boy (Gender = 0) without an IEP (Spec\_Ed = 0), the equation is:

$$\text{ELA\_Grade}' = 2.983 - 0.212(0) - 0.870(0) + 0.804(0)$$

or:

$$\text{ELA\_Grade}' = 2.983 - 0 - 0 + 0$$

<sup>15</sup>If we were predicting the standardized ELA\_Grades—which we're not—we would use the values from the Standardized Coefficients Beta column.

or simply:

$$\text{ELA\_Grade}' = 2.983.$$

So, since we coded our variables as dummy variables, then the (Constant) coefficient is the estimated ELA\_Grade score for boys without IEPs. Whatever condition in a set of data that has all os for all dummy variables is called the **reference group**: It is the group against which all effects are compared.

If we wanted to estimate the ELA\_Grade score for girls (Gender = 1) without an IEP (Spec\_Ed = 0), the equation is:

$$\text{ELA\_Grade}' = 2.983 - 0.212(1) - 0.870(0) + 0.804(0)$$

or:

$$\text{ELA\_Grade}' = 2.983 - 0.212 - 0 + 0$$

or:

$$\text{ELA\_Grade}' = 2.771.$$

It is unexpected that girls would have an estimated lower ELA grade than boys, but we also know from the bar graphs, scatterplot, and analyses that gender in fact has a rather weak effect: This estimated score is likely not strongly predictive (accurate) for any particular case.

IEP status, however, was more predictive, having an  $\eta^2 = 0.129$ , compared to gender's  $\eta^2 = 0.25$ . The estimated grade for a boy (Gender = 0) *with* an IEP (Spec\_Ed = 1) is:

$$\text{ELA\_Grade}' = 2.983 - 0.212(0) - 0.870(1) + 0.804(0)$$

$$\text{ELA\_Grade}' = 2.983 - 0 - 0.870 + 0$$

$$\text{ELA\_Grade}' = 2.113.$$

Having an IEP had a relatively strong effect on a boy's ELA grade.

The effect of an IEP on a girl's grade must take into account not only that she's a girl and that she has an IEP, but also the interaction effect of being a girl with an IEP:

$$\text{ELA\_Grade}' = 2.983 - 0.212(1) - 0.870(1) + 0.804(1)$$

$$\text{ELA\_Grade}' = 2.983 - 0.212 - 0.870 + 0.804$$

$$\text{ELA\_Grade}' = 2.705.$$

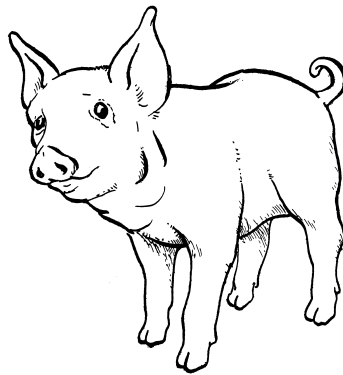
We knew from our initial investigations into the correlations between these variables that gender and IEP status were themselves related, and this is where that is represented in a linear model.

So, to summarize how the dummy variables here are set to work to create an estimated ELA grade (and to change the equation notation a bit):

Table 10.1: Example Interpretation of Dummy-Coded Variables

Boy without IEP (Reference group):	$\text{ELA\_Grade}' = b_{\text{Constant}}$
Girl without an IEP:	$\text{ELA\_Grade}' = b_{\text{Constant}} + b_{\text{Gender}}$
Boy with IEP:	$\text{ELA\_Grade}' = b_{\text{Constant}} + b_{\text{Spec\_Ed}}$
Girl with IEP:	$\text{ELA\_Grade}' = b_{\text{Constant}} + b_{\text{Gender}} + b_{\text{Spec\_Ed}} + b_{\text{Gender\_IEP\_Interaction}}$

The  $b$ -weights for the terms in each model determine the predicted score for that condition.





## Chapter 11

# Longitudinal Analyses: Why and How to Conduct Multilevel Linear Modeling

### 11.1 Overview

After reviewing a few common ways to analyze longitudinal data, this chapter then focuses on conducting a rather sophisticated approach, variously called hierarchical linear regression, multilevel modeling, or mixed models. These models can be used to analyze more than just longitudinal data, but that will be our focus here.

As is usual for these chapters on using software, I will try to present two things related to the relevant analyses. This time, the first (in Section 11.3) will be a bit more “down-and-dirty” data preparation using both a spreadsheet program and SPSS. The second (in Section 11.11) will be using SPSS to review these data and to conduct a multilevel model (MLM) of change on them.

### 11.2 Comparison of Analyses of Longitudinal Data

There are four ways that time-dependent data are typically analyzed:

#### 11.2.1 Explanations of Longitudinal Analyses

##### 11.2.1.1 Pre-Post Difference Scores

Along with repeated-measures ANOVAs, analyzing pre-post difference scores is probably the most common method of analyzing changes over time.

Of course, pre-post difference score analyses require only two waves of data collected before and after some event. One first usually subtracts a given participant's<sup>1</sup> pretest score from their

---

<sup>1</sup>Although I'll be talking about “participants” that we're studying over time, we could just as easily be studying something else that has different waves of data attached to it, such as the number of falls in a given unit over the course of a week.

posttest score to compute this difference. These difference scores are then compared between groups using, e.g., *t*-tests to see whether the mean difference between groups differs significantly.

### Advantages

They are easy to compute and intuitively easy to interpret.

### Disadvantages

Otherwise, they suck. First, any measure that is vulnerable to ceiling or floor effects will suffer from a bias in difference scores: Those who start and/or finish near the limit of the scoring range risks being affected by score restrictions: any values outside of the measurable range will be restricted<sup>2</sup>. Therefore, it is harder to detect a difference among those who either start or end the study with extreme scores—those we may well be most interested in.

Second, difference scores use only part of the information in the data. Not only can they not test for range restrictions (or other biases in the measurements of any of the waves), but they also omit any of the information available in the actual *measurement* of the phenomenon at each level; they are, after all, only the difference in the scores—not the scores themselves. Data are expensive and difference scores are wasteful.

Outside of any decisions about ethics and sophistication of analyses is the simple fact that difference scores are more susceptible to both Type 1 and 2 errors (and often [Type VII errors](#)). Since difference scores omit a large a large part of the measurement that contains the true, underlying score (e.g., a person's actual confidence in their ability to treat an illness—not just how confident they say they are), they are more likely to miss a true effect (and thus commit a Type 1 error). Since difference scores (as explained further just below) are more susceptible to error, we may also believe we've found an effect when in fact it was just a random difference due to chance alone.

### Other Problems with Interpreting only Two Waves of Data

It is worth also raising another point—not about pre-post difference scores, but about using only two waves of data to test changes over time. As argued by Singer and Willett (2003, [p. 10](#)), the interpretation difference between only two waves of data—such as pre- & posttest scores—is difficult to disentangle from error or biases in the responses. The old adage (if that's what it is) that it takes three points to test a line applies here: That third point of data can help establish a trend in the time-varying data.

Now, as one who has rather frequently used only two waves of data, I have to let my [cognitive dissonance](#) have its say. I don't disagree with Singer and Willett's argument that it's hard to tell if the difference between two waves is chance—or even a bias in one or both waves—but I don't see how this necessarily differs from any comparisons between two groups that differ in some continuous measure.

I do agree, though, that the answer to the problem of having only two waves is to have more data. I also agree that those additional data are most efficient if they are taken at other points in time—especially data measured farther out from the other waves<sup>3</sup>. However, I honestly do

<sup>2</sup>Variables like this—that can have meaningful values outside of the that our instruments can measure—are called “censored.”

<sup>3</sup>Singer and Willett explain this well in the context of obtaining more precise (defined as an efficient measure that quickly hones in on the true, underlying score). There, they note that simply increasing the variability of when (in time) measures are taken helps determine the rate (e.g., slope) of change. An example of why this is true: Imagine I want to draw a line on a wall that is parallel to the floor. I can do this by using a ruler to measure the same distance from the floor,

believe that problems of disentangling differences between two waves or between two groups can be addressed with more data—especially data that has a lot of “good” variance to it.

### 11.2.1.2 ANCOVA with a Pretest Covariate

Often when we take pre- and posttest measures, what we’re really interested in is whether groups differ in their posttest scores; we’re not interested in whether they differ at pretest—in fact, we hope they *don’t* differ at pretest. Of course, for interesting and uninteresting reasons alike, they may differ at pretest.

Pre-post difference scores attempt to account for any group differences at pretest by eliminating the information available about participant’s pretest scores. There is an other strategy, though: partial out the pretest scores from tests of differences between posttest scores.

To do this, we conduct an ANOVA in which, e.g., the groups are the IV (predictor) and the posttest score is the DV (Outcome), thus testing for group differences in the posttest scores. However, we can’t defensibly just do that since it may well be that part of an individual’s posttest score is determined by that person’s pretest score<sup>4</sup>. So, we also include pretest scores as a covariate in our model—making our ANOVA into an ANCOVA and thus effectively accounting for any effect of one’s pretest score on the posttest score.

#### Advantages

ANCOVAs with pretest covariates accurately and efficiently account for relationships between pre- and posttest scores, letting any group differences in posttests be isolated from effects of the pretest. Essentially, this type of ANCOVA lets us use math to take care of any problems with our baseline that our study design could not.

Adding pretest data as a term in our model not only isolates its variance from other terms, it also lets us tests that pretest term. A significant main effect for the pretest term would indicate that pretest scores do indeed significantly predict posttest scores. A significant group  $\times$  pretest interaction would mean that pretest scores differentially affected the groups’ outcomes (whether or not they started with differences in the pretest levels).

Indeed, to the best of my knowledge, ANCOVAs with pretest covariates are the best option outside of MLMs to test for group differences at posttest.

#### Disadvantages

Of course, these models can only be used to test just that: group differences at posttest.

If one has more than two posttest waves, one could combine this approach with the next one and create a repeated-measures ANCOVA. This method could work, but suffers from all of the disadvantages of a repeated-measures ANOVA while also removing analyses of differences from baseline/pretest that are often made in repeated-measures ANOVAs.

Other disadvantages of this approach are related to the needs of ANOVAs. ANOVAs, of course, assume that data—and error—are homoscedastic. However, this is often not the case for time-varying measures; often people who start at a similar place take different paths over time. This

---

make marks on the wall at each measurement, and then lay the rule flat along those marks on the wall to draw my line. I will have the most luck in actually making a parallel line of I make my marks at different places along the wall—and especially if those places I mark are far apart.

<sup>4</sup>Sure, maybe because of a ceiling or floor effect. But even more generally, people who, e.g., start with lower scores will also be more likely to end with lower scores.

increase in variance—both in true scores and error—is not handled well by ANOVAs. Like with having only two waves of data, ANOVAs aren't designed to account for this extra variance, so it can easily either eclipse a true effect or lead us to think we have an effect when we don't.

### 11.2.1.3 Repeated-Measures ANOVA

#### Basic Concepts

Strictly speaking, a repeated-measures ANOVA per se is *only* tests differences within participants across different times. So, if we first administer one treatment to a patient and then administer a different treatment to that same patient (measuring outcomes each time), then we are conducting a repeated-measures ANOVA.

Of course, we often have more than one group of participants, and want to compare outcomes about them at different times. This is more accurately called a two-way (or two-factor) repeated-measures ANOVAs<sup>5</sup>, but they are often simply referred to as a repeated-measures ANOVA without the additional clarification about there also being non-repeated factors in the model, too.

Like any other ANOVA, a repeated-measures ANOVA conducts an omnibus test for each factor in the model to see if there are any significant differences between the levels somewhere without testing where; specific comparisons between each of the levels are detected, e.g., with post hoc comparisons<sup>6</sup>. The difference, here, of course, is that a set of those comparisons are for different points in time that are nested within each participant.

#### An Example

I think a pre-post design with an experimental and control group serves as a great example to concretize these concepts. The executive functioning data don't work so well for this example, so let's instead use a different one. Let's say we have a health literacy program, one goal of which is to improve patients' confidence in asking probing questions with their health care provider about their condition. We measure this confidence (through some self-report instrument) for all participants before experimental-group participants complete the program (control-group participants complete an unrelated program during this time); we then re-administer this confidence instrument after the program is done, getting pre- and posttest scores from both the experimental-group participants and control-group participants.

In this design, we have two factors: the treatment—whether the participants completed the health literacy program—and time—whether measurements were made at pre- or posttest. So, in our model, we can include a main effect term for treatment; this will test whether there is an overall difference between confidence scores between the experimental and control groups. This will also use the typical treatment and error sum of squares values that one thinks of for an ANOVA.

We can also include a main effect term for time; this will test whether there is an overall change in confidence scores from pretest to posttest. The sums of squares will be partitioned a bit differently since we will be looking at differences within each participant<sup>7</sup>.

<sup>5</sup>It's also called a "two-way (or two-factor) ANOVA with repeated measures." If there are more factors in the model, it would be called a three-way, four-way, etc. ANOVA with repeated measures.

<sup>6</sup>I've avoided mentioning this pretty much every time I mention post hoc comparisons. Even within the realm of level comparisons, there are alternatives, most notably planned comparisons, which are conducted instead of an omnibus ANOVA followed by post hoc analyses. Planned comparisons are preferred if one knows ahead of time (i.e., they're planned) which specific subset of comparisons between levels one wants to conduct. One then uses, e.g., *t*-tests to compute those specific tests and no other inferential tests for those data. Planned comparisons tend to be more powerful, but you need some discipline (and, of course, specific questions you want answered) to use them. They do tend to get short shrift, though.

<sup>7</sup>[This website](#) has a pretty simple and clear explanation of the ways the sums of squares are computed.

Most importantly, we will include a treatment  $\times$  time interaction term. This will test whether any differences in pre-post scores are themselves different between the groups<sup>8</sup>. With a significant interaction, we could then use post hoc analyses to see, e.g., if the groups significantly differed at posttest but not at pretest and if experimental-group participants' posttest scores differed from their own pretest scores, but those of the control group did not. (We could also add directionality constraints, seeing if the experimental group's mean posttest scores were higher than the control group's; this would allow us to use one-tailed significance tests and increase our power for that test.)

So, at its heart, a repeated-measures ANOVA tests differences in levels. It's just that some of those levels are within participant, so—under the hood—the sums of squares are computed differently for with-participant (time-varying) terms.

### Model Assumptions

Repeated-measures ANOVAs must meet the same assumptions of non-longitudinal ANOVAs—mainly that data and error are normally distributed and independent. However, because repeated-measures ANOVAs span multiple waves should also meet extensions of those two assumptions. First, not only should each variable be defensibly similar to a normal distribution, but the relationships between the variables should also approximate normality; this manifests as what's called “multivariate normality,” and can be investigated by looking at the distribution of difference scores between variables (and especially the difference scores between waves for a given variable): If the difference scores are reasonably normal, then your data are more or less multivariate normal.

### The Sphericity Assumption

Second, the data should display what's called “sphericity,” which is essentially homogeneity (and independence) of variances across the waves<sup>9</sup>. If the data do not show enough sphericity, then we run the risk of under-estimating our ability to detect significant differences: The larger error variance in some waves may be hide a real effect between other waves.

It's not unusual for longitudinal data to contain non-ignorable departures from sphericity, so it is common to test for these when conducting repeated-measures ANOVAs. In fact, SPSS does this by default, reporting the results of [Mauchly's test](#), which tests for a difference between the variances. In other words, Mauchly's test looks for a significant difference a lot like a  $\chi^2$ -test does. If it finds a significant differences between the variances at different waves—if the  $p \leq .05$  for Mauchly's test—then the assumption of sphericity may be violated<sup>9</sup> and one should modify the analyses to accommodate for this violation.

To deal with violations of sphericity, one can use either the Greenhouse-Geisser or the Huynh-Feldt correction; both of these adjust the  $F$ -test's degrees of freedom as much as needed based on the extent of the departure from sphericity. The Greenhouse-Geisser correction is a bit more commonly used—and may be more accurate especially for larger departures from sphericity—but both tend to return similar results for data that are good enough to analyze. Generally, you

<sup>8</sup>If you've been reading these footnotes and thinking really deeply about 11 of this, well, first, thanks. Second, if you have, then you may have reflected that if it really is simply whether the experimental-group participants showed greater pre-post differences and did the control-group participants, then we could simply conduct a planned comparison of this—you're right! If you also reflected how this kinda looks like a t-test of pre-post differences scores like I first listed in this section...yeah, you're right.

<sup>9</sup>It's called sphericity (it's also called circularity, and a type of compound symmetry) because it posits that—with three waves—the variance in wave 1 equals the variance in wave 2 which equals the variance in wave 3, or  $\sigma_{wave1}^2 = \sigma_{wave2}^2 = \sigma_{wave3}^2$ . They create a little circle—or I guess sphere—of the variances all being equal to each other.

should be safe reporting the the Greenhouse-Geisser correction unless one of them indicates significance and the other doesn't, in which case use the Greenhouse-Geisser correction if the epsilon value in SPSS is lesser than .75 and use the Hyund-Feldt correction if epsilon is equal to or greater than .75.

Both the Greenhouse-Geisser and the Hyund-Feldt corrections are also given by default in SPSS when one chooses Analyze > General Linear Model > Repeated Measures. They appear in the Tests of Within-Subjects Effects table, where the adjusted  $p$ -values in that table (given as Sig.) can be used to test significance of the repeated variable; if this value is less than .05, then there is at least one significant difference somewhere between the waves that we can be relatively sure is real, so we can then, e.g., conduct post hoc analyses to find the difference(s).

### Advantages

Its comparisons between scores at different waves is intuitive, and many readers and reviewers are accustomed to interpreting it. If one simply wants to test differences between fixed points in time (and if the assumptions of the model hold)—like we did in our example of confidence pre- and post-participation in a health literacy program—then repeated measures work well. We can also run repeated measures analyses on multilevel (i.e., nested) data. For more about doing this in SPSS, UCLA's Institute for Digital Research & Education has a [good guide](#).

### Disadvantages

Repeated-measures ANOVAs are vulnerable to missing data, and missing data tends to be a bigger problem in longitudinal designs (believe me!). If a participant is missing data for any of the waves, then we remove their data from all of the waves; essentially missing data are handled with listwise deletion in them.

Repeated-measures ANOVAs make equal comparisons between waves—a test of a difference between, e.g., wave 1 and 2 is tested the same as a test for a difference between wave 2 and 3; given this, tests between waves may not be valid if the waves are not all equally spaced.

### Conclusion

Repeated-measures ANOVAs are good for what they do—if you can keep the study together long enough to maintain equally strong measurements and throughout it. If you can keep variability, sample sizes, and distances between waves reasonably equal—and if you can compensate for any lack of sphericity between the waves—then repeated-measures ANOVAs are a good way to test differences between waves and between groups within a given wave.

#### 11.2.1.4 Multilevel Models of Change

##### Basic Concepts

Both repeated-measures ANOVAs and multilevel models of change parcel out variance and covariance based on which terms are nested within which other terms. Since we make multiple measures of a given participant (over time), we are nesting those measurements within that individual in both types of analyses. Multilevel models (MLMs—also called hierarchical linear models, HLMs, and linear mixed models, LMMs) allow for one variable to be “nested” within another variable like a repeated measure ANOVA, but they also allow us to model time differently—and more

fully and flexibly<sup>10</sup>. In a repeated-measures ANOVA, the various waves of time are modeled as levels of a nominal variable—just like any other nominal variable in an ANOVA. MLMs model time as a continuous variable. As we discussed in class, from this continuous variable of time, MLMs compute both a slope—to measure how the outcome changes over time—and an intercept—to factor in the effect of where participants' initial levels begin and how that may affect other things in the model.

Both repeated-measures ANOVAs and MLMs parcel out variance and covariance in the model based on what level a term is at. So, both model time as being nested within the participant (what Singer and Willett call a [level 1 model](#)) and both model between-participants effects at a different level (Singer & Willett's [level 2 model](#)). MLMs, however, cut the variances and covariances into smaller, more precise pieces. This allows us to model both relationships between terms more precisely and to account for error more specifically.

### Assumptions

MLMs typically assume that the error terms (or more specifically, the model residuals) are normally distributed with means of zero.

We also assume that the residuals are unrelated to the model's predictors and that the residuals for one term (say rate of change within a participant) are unrelated to residuals in another term (say a between-participant treatment term), but these are usually taken care of by correctly adding both terms to the model.

Note that we do not assume [sphericity](#). In fact, MLMs are designed with the expectation that a participant's score at one point in time will be related to that person's scores at other points in time—and that a participant's residuals at one point in time will be related to residuals at other points in time. They also are robust against differences in variance across time, i.e., that scores may well spread out (or contract) across time.

We do still want to try to minimize our error terms. We also still want to ensure that we correctly model ways in which people or situations are similar—such as being treated at the same hospital or being in the same research group.

So, for MLMs, it is good practice to look at the residuals for the various terms in the model to see if (1) they appear to be normally distributed, (2) they are not skewed, (3) and that they do not correlated with any predictors or criteria in the model. Of course, if any of the residuals *are* correlated with any of the term (predictor or criteria—or perhaps even other residuals) then good! There is something else interesting going on in your studies that may produce novel and fruitful insights in future research. But for now, any residuals that are egregiously non-normal or correlated with model terms would simply encourage one to try out different combinations of terms to look for a better-fitting model, and then to report, say, this in the Results and then conjecture about it in the Discussion before putting that thought into an induced coma until you can revisit it more systematically.

### Advantages

Modeling time as a continuous variable has several implications for the model. First, we don't need to have the waves equally separated from each other. It's fine to have the waves at unequal intervals—just make sure to measure “where” that wave is in time, e.g., the number of days that wave occurred into the study.

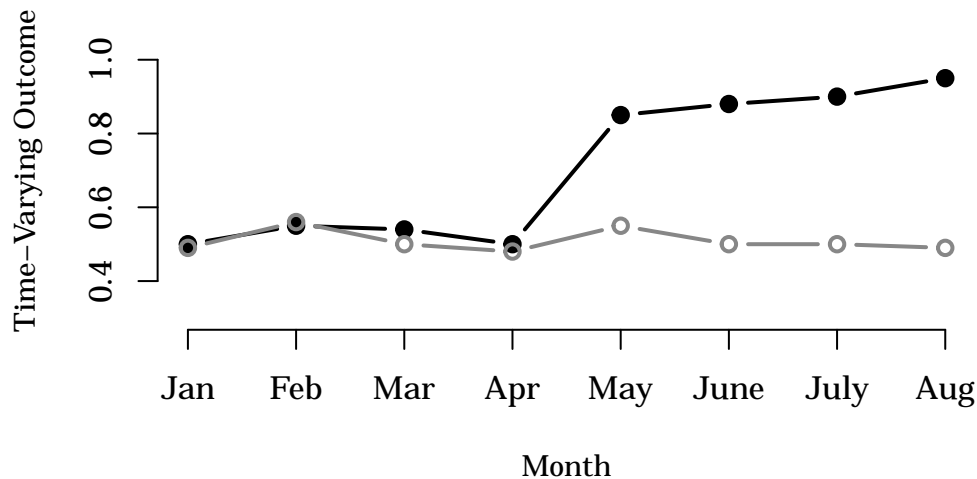
<sup>10</sup>There are other ways to handle time—and even ones more flexible than MLMs. Latent growth curve (LGC) modeling is one example that allows even finer tests of change over time, e.g., that different participants show different growth curve (say, some showing linear growth while others show exponential.)

Second, the participants don't need to have their waves at the same time. Because the slope is estimated within each participant, measuring outcomes at different times for different participants will not adversely affect the analyses. Of course, it may affect our interpretations if we're trying to infer events happening at one time for one participant with events happening at a very different time for another participant, but that's more a matter of interpretation, not computation.

Third, I don't need to have the same number of waves for each patient. Having missing time data in a repeated-measures ANOVA removes all other data for that participant. For a MLM, having fewer waves for a given participant simply means that estimates for that participant's slope<sup>11</sup> are less powerful since there is a smaller sample size.

### Disdvantages

Not all time-varying events change in a simple, linear fashion. For example, although simply plotting a regression line through each these two groups would show changes over time (i.e., have different slopes):



it would miss the important detail that it's only in May that the groups start to differ. (Repeated measures ANOVA would be a better choice for here, assuming the waves of time were evenly spaced, error is homoskedastic, etc.)

Another disadvantage is that MLMs do take more sophisticated understanding to conduct—and to understand as a reader. Sometimes, the best analysis is the one your readers understand best.

#### 11.2.1.5 Regression Discontinuity in Time

Regression discontinuity in time (RDiT)....

### Further Resources

- RDiT versus other methods, including comparisons of it with traditional regression discontinuity designs

<sup>11</sup>As well as for estimating the intercept if the intercept is set to be something other than one of the waves, e.g., if you don't/can't assume the first wave of data is the intercept for the participants.

- Hausman, C., & Rapson, D. S. (2018). [Regression discontinuity in time: Considerations for empirical applications](https://doi.org/10.1146/annurev-resource-121517-033306). *Annual Review of Resource Economics*, 10(1), 533–552. <https://doi.org/10.1146/annurev-resource-121517-033306>
- Hausman, C., & Rapson, D. S. (2017). [Regression discontinuity in time: Considerations for empirical applications](https://doi.org/10.3386/w23602). National Bureau of Economic Research Working Paper Series, No. 23602. <https://doi.org/10.3386/w23602>.

\* This version is a little longer and more detailed than the 2018 version published in the *Annual Review of Resource Economics*, just above.

## 11.3 Data Preparation and Manipulation

The data we will be those from which the data we used previously (viz., EF\_Slope\_Data.sav) were derived. The data we'll use here also contain more participants from the school.

More importantly, they are also less prepared for the analyses here; they are both spread out over multiple files and are not all in the “long” format required for MLMs. It is not uncommon for data to come in the “wide” format that are not conducive to MLMs, so covering what wide and long formats are and how to change data between these formats is nearly necessary here.

In addition, data may come in multiple files like I'm presenting here; it's how I store these data in part because I get them from different sources. In addition, the sets of data related to this line of research already include over 5 million data points, so I rarely evoke them all together (and hardly have the RAM for it even if I wanted to). Instead, I merge data stored in different sets as needed. Although this means keeping track of different sets, I've found it helps keep the data clean since each set is simpler, and—as I hope you'll see—allows me to store them in ways more appropriate for those particular data.

We'll be combining three sets of data here:

- demographics.sav, a matrix of data that includes gender, race/ethnicity, economic distress, English language learner status, birth date, and IEP status;
- ef.sav, a “long-”formatted data set that contains information about executive functioning scores; and
- discipline.sav, a “wide-”formatted data set that contains the number of discipline incidents of nearly any kind (ranging from a uniform infraction to assault).

*Remember that, although anonymized, this is information about real people—teens who are struggling to grow and succeed—so please treat these data kindly and confidentially.*

Let's first look at each of these data sets before merging them together.

## 11.4 Understanding the demographics.sav Data

This set of data is most like those you're accustomed to: The id identifying each adolescent appears in the first column; also in each row are data related to that particular teen. There is one row for each teen, and each column contains data related to aspects of that teen's demographics. The first of these is a binary conception of gender, where—again—males are coded as 0 and females as 1. The next column contains a text description of that student's race/ethnicity (“African-American,” “Asian-American,” etc.). The next five rows recode race/ethnicity into a set of dummy variables;

note that I kept a dummy variable for all races/ethnicities—I didn’t exclude one ethnicity to serve as the default (reference) comparison (“Multiracial” was coded out into the races/ethnicities that comprise that group).

After the dummy race/ethnicity variables are a set of other dummy variables. These are for economic distress, whether the teen is an English language learner, and whether the teen has an IEP—all variables you’re used to working with by now.

The final variable, `julian.birth`, is in fact the teen’s birthday. It is expressed in the Julian system, and is simply the number of consecutive days since a standardized, arbitrary date<sup>12</sup>. Although the dates can’t be read without using a [conversion](#), it is useful here since we will be looking at longitudinal data for which measuring things on a scale of days can be useful<sup>13</sup>.

## 11.5 Understanding the `discipline.sav` Data

This file also contains `id` in the first row, so that rows with the same `id` in the `demographics.sav` and `discipline.sav` files relate to the same teen. The next three columns in the `discipline.sav` data set give the number of times each year that that student was reported to the main office for disciplinary actions related some sort of rule violation. So, the second column, `AY_20162017_Incidents`, is the number of times a student was disciplined for violations during the 2016-2017 academic year.

(Note that I did not separate out the severity of the infraction or the discipline. Therefore, the discipline could have been for something rather small, like talking back to a teacher, or fairly large, like getting in a physical fight. However, most very-small infractions are not relayed to the administration, and very serious incidents result not in discipline but in, e.g. expulsion. Therefore, the infractions here tend to be middling ones, like multiple absences or verbal altercations with other students.)

The data in this set are presented as one often sees for chronological data: Each row is a different person/case, and each column a time-related datum for that moment in time and for that person/case. One advantage of data presented in this “wide” format is that it is rather easy to see for whom there are data and to get a sense of patterns across the time moments for a given person. We see, for example, that the fourth student from the top, 219983160, had 3 incidents in academic year (AY) 2016-2017, 5 in the next year, and then none in the final year: Perhaps after a couple rougher years of adjustment, this student is now catching their stride. It is also the form data should be in for one to compute a repeated-measures ANOVA.

As we will see, though, there are analysis-related advantages to re-arranging the data from this “wide” format to a “long” one.

## 11.6 Understanding the `ef.sav` Data

The `ef.sav` file contains data related to each student’s executive functioning. The right-most columns are `metacog.index.tchr`, `beh.reg.index.tchr`, and `global.exec.comp.tchr`: meta-cognitive,

<sup>12</sup>It is the number of days since January 1, 4713 BCE, presuming use of the modern Western calendar to count back that far. It was developed by an historian, Joseph Scaliger (whence the date range), and named after his father, whence the important-sounding name. Note, however, that the dates here are not computed along the lines set out by Scaliger, but a derivative method used by a standard package in the R programming language.

<sup>13</sup>To finally create a footnote of actual value: In the early stages of this line of research, I didn’t limit myself to using Julian age, but also, e.g., simply used the calendar year or—at most—semester. However, those measures of time proved not to be precise enough to track the rather subtle effects we investigate here. Only when I used Julian dates—i.e., only when I measured time as precisely as I could—did I reliably find interesting insights. I guess I shouldn’t have been surprised that analyses of times-varying effects benefited from good measures of time.

behavioral-regulation, and the composite global executive composite functioning scores, respectively, as measured by one of that student's teachers; and `metacog.index.sr`, `beh.reg.index.sr`, and `global.exec.comp.sr`, scores on similar domains as self-reported by that same student. The meta-cognitive, behavioral-regulation, and global executive composite scores are all obtained from the same instrument, so these three scores are obtained on the same day for each occasion.

Before these columns containing EF-related scores are two columns related to time. The column right after `id` is a good place to start to understand these. In the first row of data, we see the student-reported EF scores for student 2 for the 2010-2011 academic year. In the second row, we see the EF scores *also* for student 2 for the 2011-2012 academic year.

You can therefore see that I have organized the data in this set so that each row includes any and all data collected at a given moment in time. Data were collected on student 2 on two occasions: They completed the instrument (the BRIEF-SR) to generate EF-related data during AY 2010-2011, and we were able to get STEM<sup>14</sup> grades for student 2 in AY 2011-2012.

Data that are presented in this format—with each moment in time for each person given in a different row—are called “long” data; that this data set spans nearly 8,000 rows explains why it's called that.

The other time-related column, `wave`, contains the Julian-style date on which those data were collected. This number is even less intuitive than the `julian.birth` field. In my own set of these data, I call this field `julian.wave.semester.centered`, which belies some of why these values are so abstruse: It is a day near the end of a given semester, that part isn't so hard to understand. It is a Julian date, but obviously not the number of days since 4713 BCE; instead they have been “centered” so that we are counting the number of days since the school opened, 304 days before we first collected data on any students. I also use the convention suggested by Singer and Willett (2003) and others to call each moment in time when data are collected a “wave.”

## 11.7 Restructuring and Merging the Data Sets

In order to merge these three sets of data in preparation for a multilevel model of change, we must convert the wide `discipline.sav` set to a long format and then match each set by `id`. Since the discipline data and the EF/grade data both have time-varying data within the teens, we will need to merge them both by `id` and by a `wave` (i.e., time) term.

Note that I have included in BlackBoard the final data set that is created after the restructuring and mergers for your convenience. Nonetheless, it may be instructive to go through the steps that produced it.

## 11.8 Wide-to-Long and Long-to-Wide Data Restructuring

Converting data from wide to long and from long to wide can be relatively easily done in SPSS. Note, however, that since SPSS uses several, sequential dialogues to prepare the final restructuring, we may not always know how it will turn out until it's over and too late to Undo; therefore, I typically save a backup of my data before doing something like this. Here we won't since I've already tested it, but I do suggest both doing that and always having an original version of all your data that you leave pristine and instead only modify a duplicate of that original data. (I also set the file name for that original version to, e.g., `discipline_ORIGINAL.sav` so it's clear to me that I mustn't mess with that file—and only make copies of it and work with those copies.)

<sup>14</sup>“STEM” is the rather aspirational acronym for courses related to science, technology, engineering, and math.

### 11.8.1 Wide-to-Long Data Transformation

1. With either tab of the Data Editor window focused on the discipline.sav data set,
2. Click on Data > Restructure
3. Within the Restructure dialogue, select Restructure selected variables into cases and then click Next.
4. In the next Variable to Cases: Number of Variable Groups dialogue, Under How many variable groups do you want to restructure?, select One (for example, w1, w2, and w3) and click Next again. (By selecting More than one, we could convert several different sets of time-varying variables at once)
5. In the next Variable to Cases: Select Variables dialogue, under Case Group Identification, select Use selected variable and enter id in the Variable: field that appears.
6. Under Variables to be transposed, select trans1 in the Target Variable field and replace with a new name, e.g., disc\_incidents.
7. Select AY\_20162017\_Incidents, AY\_20162017\_Incidents, and AY\_20162017\_Incidents and move them to the field under Target Variable. and click Next.
8. In the next Variable to Cases: Create Index Variables dialogue, under How many index variables do you want to create?, select One and click Next.
9. In the Variable to Cases: Create One Index Variable dialogue, under What kind of index values?, select Variable names. Under Edit the index variable name and label, replace index1 with ay, leave Label empty (to avoid trouble with merging data sets later on), and click Next.
10. In the Variables to Cases: Options dialogue, we could set how to handle other variables in the data set that were not being restructured into long format. We don't have any, though, so simply click Next.
11. In the Finish dialogue, select to Restructure the data now and click Finish.
12. SPSS will return an error message saying Sets from the original data will still be in use in the restructured data. Open the Use Sets dialog [sic] to adjust the sets in use. This is letting you know that variables in the pre-restructuring data set (viz., id) will still be used in the newly-restructured data. This is O.K. here, so click OK.

The newly-structured data set now has three columns: one for id, one for ay, and one for disc\_incidents. We restructured these data as a first step in merging them with the other, long-form data (viz., ef.sav); in the ef.sav data set, the ay variable has values of, e.g., 2010-2011, not AY\_20172018\_Incidents as it does here. To allow us to easily merge these two sets of data, we will now change the values in discipline.sav's ay variable to be like those in ef.sav.

1. Click on Transform > Recode into Same Variable
2. Select ay to be added to the Variables field, and click Old and New Values
3. In the Value: field under Old Value; type (or copy and paste from here) AY\_20162017\_Incidents. In the Value: field under New Value, type 2016-2017, and then click Add.
4. Repeat this for transforming AY\_20172018\_Incidents to 2017-2018 and AY\_20182019\_Incidents to 2018-2019, click Add after each.
5. Click Continue when these three transformations have loaded into the Old --> New field; Click OK when you return to the original dialogue window for this transformation.
6. Save these data; we're done with them for the moment.

## 11.9 Long-to-Wide Data Restructuring

If you want to change data from a long format to a wide one:

1. Since we won't be using these wide data, we will first make a copy of them to play with. So, with the restructured discipline.sav data focused in either tab of the Data Editor window, click on Data > Copy Dataset. This will create a new .sav file called something like Untitled2.sav. We will work with this data set for the rest of this example.
2. With this new set focused, click on Data > Restructure
3. In the dialogue that opens, select to Restructure selected cases into variables, before clicking Next.
4. Add id to the Identifier Variable(s) field and ay to the Index Variable(s) field; click Next.
5. Under Sorting the current data? dialogue in the Cases to Variable: Sorting Data dialogue that next appears, select Yes -- data will be sorted by the identifier and index variables, and then click Next.
6. In the Cases to Variable: Options dialogue, select Group by original variable and click Finish.
7. We again get the same Use Sets error, which again is simply a warning and can be ignored here.

The incidents will now be put back into wide format, with AY 2016-2017 incidents likely called v1, etc. *Voilà*, we can now simply close and not save this Untitled2.sav file and move on.

## 11.10 One-to-Many Match Merge

The ef.sav file contains the most rows for each student, and nearly all of the ones that the discipline.sav file would populate further, so we will merge discipline.sav into ef.sav.

1. With ef.sav focused, click on Data > Merge Files > Add Variables...
2. In the dialogue that opens, select discipline.sav under An open dataset (or select it via As external SPSS Statistics data file if it isn't open).
3. In the Merge Method tab in the next dialogue:
  1. Select One-to-one merge based on key values. We will match cases based on both id and ay, but here each id and ay combination is unique, so it's a one-to-one merge.
  2. In the area below that, make sure to select Sort files by key values before merging. (Both files should be sorted already as need be, but it's always good to ensure they are since the match merge in SPSS requires this to work correctly.)
  3. Make sure that *both* id and ay are populating the Key Variables field further down; we will be merging based on both a given student's id and by the academic year for the given number of incidents.
4. In the Variables tab, there are no variables to exclude, so click OK.

This should work fine, but let's just check it to be sure:

1. In the ef.sav file, right-click on the disc\_incidents variable and choose Descriptive Statistics from the drop-down menu. Do this for the disc\_incidents variable in the discipline.sav file as well.
2. In the Output window, ensure that both instances of the disc\_incidents variable have 1364 valid cases (of course, the number of invalid ones will differ) with a mean of 5.99 incidents. Checking the rest of the descriptives further supports that this process proceeded correctly.

We will now merge the demographics.sav file in with this merge set of data. The demographics.sav data set has one row for each adolescent while the ef.sav set often has more than one row for each adolescent, so we will add the information from the demographics.sav file to more than one row—a one-to-many merge.

1. With the merged `ef.sav` file focused to be the active data set, we again access the requisite dialogue via `Data > Merge Files > Add Variables...` and now select `demographics.sav` from the appropriate source to merge into the current data set before clicking `Continue`. (First, though, note which DataSet `demographics.sav` is labeled as since this can help later when figuring out which is which; the exact number, e.g., DataSet3 or whatever, will depend on the order they've all been opened<sup>15</sup>.)
2. In the `Add Variables from DataSetX` dialogue, under the `Merge Method` tab, choose for this to be a One-to-many merge based on key values since we'll want to load the respective demographic variables into each row for a given student, regardless of which which it is.
  1. Under `Select Lookup Table`, select whichever DataSet is the `demographics.sav` file. This should be the DataSet noted in the dialogue title, `Add Variables from DataSetX`, and is the one that could have been noted in step 1 of this merge. In addition, if you indeed had `ef.sav` focused to be the active data set, then `demographics.sav` will be the DataSet not asterisked to be the active one.
  2. Again ensure to `Sort files by key values` before merging.
  3. Since we'll want to demographic values to populate each row for a given student, we only want `id` alone to be in the `Key Variables:` field.
3. Under the `Variables` tab of the `Add Variables from DataSetX` dialogue, we can keep all variables again. Ensure here, though, that `id` is listed in the `Key Variables:` field.

If all went well, then the `ef.sav` file should now have all of the data from the three data sets in long format. Let's also now click on `File > Save As...` and save this as, e.g., `all_data_long.sav` (or even `Data > Copy Dataset`).

With the data in this form, we can evaluate multilevel models of change fit to these data. First, though, let's consider other ways we could analyze longitudinal data (in wide and/or long forms) and compare their respective advantages and disadvantages.

## 11.11 Conducting a Multilevel Model of Change

I hope I've convinced you that MLMs are worth giving a try. We can use SPSS's GUI for most of this, but I also hope you'll see the advantages of using its syntax. I wouldn't expect you to use the syntax all of the time, but there are times when it's useful—and times when it's needed since many of SPSS's vast number of commands and subcommands are not accessible through the GUI. So, if you are going to expand your repertoire of analyses to improve your opportunities (and if you're going to use SPSS), then you'll likely eventually want to venture into its syntax.

This also makes my explanations of how to conduct the analyses different since there's little need to go step-by-step through them. Instead, I will present the syntax and explain what the different parts mean.

But first, let's get some familiarity with using SPSS's syntax to do anything.


### 11.11.1 Introduction to SPSS's Syntax

As noted in Section 17.2.3.2, as of version 28, the syntax SPSS used to generate output is no longer presented by default in the output, but `Edit > Options > Viewer > Display commands` in the log

<sup>15</sup>For match merges like this, it's arguably easier to have only the data set to which you want to add files open and to access the closed data set from `An external SPSS Statistics data file` field in the `Add Variables to X` dialogue since you needn't worry about which open set is which, but I wanted to show you the more involved way so you know how to do that. The other way can be sufficiently covered in a simple footnote.



will let it be subsequently presented there. With the syntax is presented in the output<sup>16</sup>, we can copy and paste that syntax right into the Syntax window to rerun or to modify and then run.

Let's briefly go over the pieces of SPSS syntax through an example doing just that.

1. We will be using the merged, long-form data set for the rest of this chapter, so close the other sets of data.
2. With `all_data_long.sav` open, click on Analyze > Descriptive Statistics > Descriptives...
3. Add `stem_grades` and all of the executive functioning scores to the Variables(s) field and click OK
4. Right above the tables of Descriptives in the Output window is the syntax SPSS used to compute those descriptive stats. You can see in the thin navigation pane to the left of the output, the syntax is labeled as Log; clicking on the icon just next to Log (  ) will take you to that syntax. We can double-left click into that box of syntax in the output window to copy the syntax; we could then click File > New > Syntax to open the Syntax window, and paste that syntax into it—but don't do this.
5. Instead, we can also skip all of that: Again click on Analyze > Descriptive Statistics > Descriptives... The same set of variables should still be in the Variable(s) field (and all options, etc. the same, too). With the same variables, etc. chosen, instead of clicking on OK, click on the Paste button right beside of it. This will open up the Syntax Editor window with the syntax already pasted into it (or appended to any syntax we already had in that window).
6. For the Syntax Editor window, we can run this command, but let's first look it over
  1. SPSS syntax usually start with a procedure, which is essentially the verb of the syntax "sentence:" It tells SPSS what sort of analysis or command you will be executing. This procedure is presented in blue text in the Syntax Editor window. The syntax we pasted in has two procedures: a `DATASET ACTIVATE`<sup>17</sup> procedure which tells SPSS which data set we wanted to run the subsequent procedure(s) on and a `DESCRIPTIVES` procedures that is followed by a bunch of stuff I'll explain next.
  2. Immediately following the `DESCRIPTIVES` procedure name is `VARIABLES` in red. This is what SPSS calls an option keyword. As you can tell, in this case, it indicates which variables the `DESCRIPTIVES` procedure should be run on. Immediately after this is an equals sign followed by the variables in our data set that this `VARIABLES` option is defining.
  3. On the next line, we see `/STATISTICS` in green. This is called a "statement," or subcommand of `DESCRIPTIVES`. In any case, these subcommands are usually specific or additional analyses to be run, and they are always preceded by a forward slash. Here, it indicates which of types of descriptive stats to run—each statistic itself being an option.
  4. One more thing to note is that SPSS procedures not only always start with a procedure subcommand (here, of course, that's `DESCRIPTIVES`), but they also always end with a period.
7. I'm sure that was very edifying, so let's now build on that with more, interesting insights into the syntax:
  1. We can, well, edit the syntax in the Syntax Editor. In fact, we could have typed it all out, or—as we will do later—copy and paste syntax from other files.
  2. Let us add another subcommand to the `DESCRIPTIVES` procedure. Left-click into the field in the Syntax Editor right before the forward slash in the `/STATISTICS` subcommand and hit the return key on your keyboard.

<sup>16</sup>We can also get the syntax from the [journal file](#). The location of the journal file can be found by going to Edit > Options > File Locations and looking under the Session Journal section for the Journal file field. Note that Record syntax in Journal must be selected, but it is by default.

<sup>17</sup>SPSS puts certain elements in upper or lower case to help distinguish things, but case doesn't. You can write SPSS syntax in what case you want.

3. In the empty line above the /STATISTICS subcommand, type /SAVE<sup>18</sup>.
4. This is a rather cryptic subcommand, so let's further modify the syntax. Right after the /SAVE subcommand (on the same line), type<sup>19</sup>:
  1. /\*This subcommand generates standardized variables for all the variables included in the VARIABLES option.
5. This is a comment, a piece of text that SPSS ignores but that we can use to help us understand what something means (like we did here) or why we're doing something so that we can reorient ourselves if we come back to it later. Note that this comment starts with a forward slash and then an asterisk—and that it ends with a period; for in-line comments like this, we need all three punctuation marks.
8. We could save this modified syntax if we wanted (Ctrl + S<sup>20</sup>, File > Save, or the floppy disk icon ). SPSS syntax is saved with the .sps extension, but—unlike .sav files—they can be opened by other programs, like text editors ([Notepad++](#), [Atom](#), [Kate](#), [Vim](#), [Emacs](#), etc.)
9. Of course, we can also run it. To run all of the syntax in a window, click the green “play” button  or press Ctrl + R. This will run the procedure(s) that are highlighted or the procedure in which the cursor is currently placed. We could also click Run > All to run all of the syntax in that window.

Being able to save/rerun analyses and to insert comments into syntax are perhaps the main reasons to use it for procedures that one could otherwise access from the GUI menus. As I've noted before, I recommend keeping a sort of journal<sup>21</sup> of what you did. You can do this by adding text and notes directly into the Output produced, of course, but you may also want to clean up that output to be just the analyses that mattered—not, e.g., all the exploratory things you did. One way to create a simple and replicable history of your analyses is through a commented series of syntax-only procedures.

As I noted above, an other reason to use SPSS's syntax is that not all procedures or options are available through it, including those to fully conduct MLMs. Now, some procedures herein are accessible via the GUI, but since many are not it seems awkward to switch back and forth; it also undermines the purpose of demonstrating using SPSS's syntax interface.

In addition, you needn't rely just on this chapter for help conducting MLMs in SPSS. All of the syntax I'm using based on that presented in the [companion website](#) to Singer and Willett's book. Although I modified what they give there, this means you can use the excerpts from the book I provided along with that site (and this chapter) to have what I hope is a strong foundation in MLMs—analyses I am a clear advocate of.

### 11.11.2 Overview of Analyses

We will investigate how well economic distress and other demographic variables predict the development of these teens' executive functioning. For these analyses, let us assume that the variable

<sup>18</sup>You may notice that SPSS proffers suggestions for syntax as you type. You can select the appropriate command from the proffered list by hitting Enter. In addition to this as a guide to possible options and how to type them, you can access a syntax reference sheet via Help > Command Syntax Reference.

<sup>19</sup>Or don't actually type all of that; just know that you can add comments like this, and that that is what this subcommand is doing.

<sup>20</sup>Command + S on a Mac.

<sup>21</sup>Indeed, SPSS can create just that via Edit > Options > General where you can select Record syntax in journal to have any syntax—including that accessed via the GUI—recorded to a .jnl file.

of interest to us—say the one that addresses one of our research questions/hypotheses—is economic distress and that all of the other demographic variables are simply things we want to partial out. For example, our hypotheses could be that:

Adolescents experiencing economic distress will self report fewer, initial executive-functioning-related behaviors in sixth grade than peers not experiencing this distress; those experiencing it will also self report relatively weaker subsequent gains in executive functioning throughout middle and high school.

In addition, economic distress will significantly improve predictions of executive functioning made by other, apposite demographic factors, namely gender, race/ethnicity, and whether the participant has an IEP.

Earlier in our report, we could discuss how economic distress is often associated with having an IEP and race/ethnicity but that we believe that economic distress affects executive functioning in ways that are independent of these factors. We could also discuss how gender is often found to affect the development of executive functioning, and so it should be accounted for as well (nearly as a delimitation).

Note that the above hypotheses do not posit any interactions between the factors. We proposed that the effects of economic distress on executive functioning will be at least partially independent of the effects of demographics—that the effect of economic distress on executive functioning is not only through demographics. However, we did not postulate that that one's demographics will either amplify or reduce the effect of economic distress. We indeed could also test hypotheses about whether the factors do interact; however, we won't until after we've gotten the basics straight.

I will follow the general sequence of analyses recommended by Singer and Willett beginning on [page 92](#). Therefore, first we will investigate whether there are differences in these teens' executive functioning that warrants further investigation—that there are differences in their initial levels of executive functioning. Second, we will test whether there are significant changes in these initial levels—that there is indeed a reason to investigate changes over time. Third, if both these preconditions are met, we will compute a base model that contains only the demographic variables we wish to control for. Fourth and finally, we will add economic distress to that base model to see if this final model is a better predictor of executive functioning than the base model.

### 11.11.3 Computing the Unconditional Means Model

In an ANOVA, we first run an omnibus test of a factor to see whether additional tests of it are warranted. Singer and Willett recommend taking a similar tack with MLMs; they suggest first conducting a pair of analyses to test whether we have sufficient reason to look further at the data. One analysis of the pair tests for differences in initial levels of the outcome; the other of the pair tests for changes over time. Remember that in MLMs, we retain two pieces of information about the time-varying Outcome: its intercept (or initial value) and its slope (or subsequent changes over time); this pair of tests thus test if we need to indeed include both pieces of information in our model or if instead a simpler model that excludes one of these is recommended.

Note that I have never seen member of this pair reported in published article. This may be in part my limited exposure to their uses, but it still doesn't make their reporting common. I agree that it's good practice to compute them, but it's not conventional to then also report them.

With that in mind, let's run the model and use it to further our understanding of MLMs.


1. In SPSS, click File > New > Syntax
2. Paste in the following syntax into the main field in the window that opens:

```

DATASET ACTIVATE all_ef_long.
TITLE "Unconditional Means Model, p. 92".
MIXED global.exec.comp.sr
/PRINT=SOLUTION
/METHOD=ml
/FIXED=intercept
/RANDOM intercept | SUBJECT(id) COVTYPE(un).

```

Note that this and all of the syntax below are available in the `mlm_syntax.sps` file in Black-Board.

3. Select the entire set of syntax (e.g., with Ctrl + A or by highlighting it with your mouse) and then either type Ctrl + R or click the green “play” button ().

#### 11.11.4 Interpreting the Syntax

This will generate a set of output but let’s first go through this syntax to understand better what it’s doing:

- `DATASET ACTIVATE all_ef_long.`
  - This simply ensures that the `all_ef_long.sav` data set is the “active” one that will be used for any further work. It will remain the active data set until we either run another such command targeting another data set or we click on the Data Viewer window of another data set.
- `TITLE "Unconditional Means Model, p. 92".`
  - This simply outputs a title before the results echoing what’s given in the quotes. This just helps to keep the output straight.
- `MIXED global.exec.comp.sr`
  - The `MIXED` procedure declares that we are conducting a linear mixed model, which is another name for multilevel models. All linear mixed models / MLMs contain both fixed and random effects—this “mixing” of fixed and random effects is where they get their name.
  - We also declare here what our outcome (criterion) for this model is. (We’ll add more to this line later.) Here, we’re using the global executive composite score from the BRIEF-SR, the instrument completed by the adolescents about themselves. This score includes all of the various subscores for different, specific executive functions.
- `/PRINT=SOLUTION`
  - The `/PRINT` subcommand allows us to add yet more output to that which SPSS already spews. Here, we’re asking that the parameter estimates for the model be printed. Note that since this is a simple with just the intercept (more on that in a second), there isn’t much extra printed; this will be more helpful for later models.

- /METHOD=ml
  - With the ml option, the /METHOD subcommand requests that we use full maximum likelihood estimation. As Singer and Willett [explain](#), full maximum likelihood estimation creates estimates for all of the parameters in the model—both the predictors (or “structural”) and error/residual (or “stochastic”) parts. This seems like a good idea—and is the choice one nearly always makes—but this method tends to over-estimate the certainty of our predictor estimates, making their standard errors a bit too small and thus our confidence in any effects we find for the predictors a bit too high. In other words, using full maximum likelihood estimation increases our chances of Type 1 (false positive) errors. This over-confidence in effects of predictors reduces as the sample size increases.
  - The subcommand could use the reml option instead of the ml option. reml tells SPSS to instead compute restricted maximum likelihood estimation. Restricted maximum likelihood attempts to minimize the unexplained (residual) variance—that which is consigned to what I’ve been calling the error terms, and that Singer and Willett (rightly) called the stochastic<sup>22</sup> terms. This is appropriate for tests of the error terms, but we will not be doing those here—and you will likely rarely do them. Therefore, I recommend using the ml option while also striving to ensure that your sample sizes are “big enough” (say a couple hundred) and that you are cautious with *p*-values that are just on the significant side of .05.
    - \* reml is also the default option for SPSS, so it will be used if we do not specify otherwise. Again, I suggest over-riding this default unless there is reason not to.
- /FIXED=intercept
  - The /FIXED subcommand is the first of two subcommands in which we specify the terms in our model (/RANDOM is the second). In statistical models, fixed terms are those in which all levels of that variable are present in the model. For example, if a treatment variable can only either be “experimental” or “control” (and we include both levels in our model), then that term is considered “fixed.” Binary gender, dummy variables, and most other nominal variables are fixed.
    - \* They are called “fixed” because we are only interested in the levels that are actually presented in our model—i.e., we are not trying to make assumptions about other values not presented in the model—and so our parameter estimates are unchanging, or “fixed,” to those particular values.
  - Fixed effects differ from random terms. Random terms are those in which not all levels of that variable are present. Sure, this could be, e.g., if you believe not all types of race/ethnicity are present, but in fact cases like that are typically relegated to being fixed terms anyway, the missing levels being ignored. Instead, random terms are essentially any variable that’s defensibly interval/ratio, i.e., a variable that has a continuum of values—like height or even time, both of which can be measured to an infinite number of decimals.
    - \* They are called “random” because we are assuming that the actual values we have in our data set were chosen at random from a huge pool of possible values (i.e., we sampled from a larger population of values).
    - \* What about Likert-scaled variables? One usually assumes that they are random variables unless there are, say only three or four possible levels, in which case we may want to consider treating them as a series of nominal variables. (The lesson here is to try not to have Likert responses that only have a few levels.)

<sup>22</sup>“Stochastic” in statistics denotes randomness that is indeed truly random, i.e., that indeed have no bias.

- `/RANDOM intercept | SUBJECT(id) COVTYPE(un)`.
  - Finally, the `/RANDOM` subcommand does double duty in MLMs of change. First, it includes—get this—random terms. However, there are two important options added to the `/RANDOM` subcommand after the bar (“|”): `SUBJECT` and the `COVTYPE`.
  - The `SUBJECT` option indicates which variable denotes the level of the participant. It *also* indicates what other variables are nested under participant by having that nested variable given right before the bar. Right now, this doesn’t make much sense (sorry), but it will in subsequent models we specify, so please just keep this in mind for later.
  - `COVTYPE` defines the covariance structure of the model. The covariance structure is nearly always the covariance *matrix*, so we can generally assume that when SPSS says “covariance structure,” it’s referring to the covariance matrix.
    - \* I’ve not emphasized covariance matrices in class, but they hold a central role most much of what we’ve been talking about. You will remember that the covariance matrix of a list of variables is the unstandardized correlation matrix. When we compute linear regression models like we’re doing here, what we’re in fact doing is trying to maximize the chance that the values in a covariance matrix of the parameters are the correct values for those interrelationships.
    - \* There are several types of possible structures that SPSS can use; some of these are quite useful indeed, but all are beyond the pale of our needs here: Simply use the `un` (“unstructured”) type here.

### 11.11.5 Interpreting the Results

After some information about when and from where the analyses were run, SPSS first generates an error message:

---

The covariance structure for random effect with only one level will be changed to Identity.

---

An “identity” in mathematics is a term that does not change the value of an other term. For example, if I multiply a number by 1, then it stays the same number; in this case, 1 is an identity. A matrix is an identity matrix if it would not change the values in an other matrix when those two matrices are multiplied together; in practice, an identity matrix simply has ones in the diagonal and zeros in the places off the diagonal, like this:

$$\text{A } 3 \times 3 \text{ identity matrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

There is only one intercept for each adolescent (`id`), and there are no other terms in the `/RANDOM` subcommand. If I created a covariance matrix of the values of the intercepts for the `ids`, then the diagonal would be ones—since the intercept for each adolescent shares all of its covariance with itself (just like an item is perfectly correlated with itself)—and the off-diagonal values would all be zeros—since we are assuming that the intercepts for the adolescents are all independent of each other. This would create an identity matrix.

So, all this error message is saying is that. That since we only have one `/RANDOM` subcommand term for each adolescent, we’re assuming that the intercepts are all independent of each other^[All of that just to say, yeah, you can ignore the error message..

The next table further exemplifies this. It shows that the only term in the model is the intercept, and that this creates an identity matrix (Covariance Structure) among the participants (the ids in the Subject Variables):

		Model Dimension <sup>a</sup>			
		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
Random Effects	Intercept	1	Identity	1	id
Residual				1	
Total		2		3	

a. Dependent Variable: global.exec.comp.sr.

You may well be confused that the intercept term is being considered as both a fixed and an random effect—or simply that it should even be considered as a fixed effect since global.exec.comp.sr scores can clearly take on values not measured here (and thus is indeed a random variable in this sense). Indeed, we are forcing the model to do things it's not really designed to do here<sup>^</sup>[And so starting our understanding of MLMs with an aberrant model isn't perfect.. We will only use this model for one purpose, though, and it does meet that one goal. So, please otherwise ignore this odd structure.

#### 11.11.5.1 Information Criteria

The two pieces of information worth noting are given in the next two tables. I will explain them in some detail now, but will not emphasize them in most of our analyses, instead concentrating on interpretations that should feel more familiar.

The first table presented after the Model Dimension table provides the Information Criteria for the model. Information criteria are important concepts and statistics in models. They represent the amount of information left in the data that is not well accounted for by the model. Yes, it's the same concept as the error sum of squares in an ANOVA. Similarly, we are trying to minimize this residual—this unexplained variance—so a core goal here is to minimize the amount of information left in the data unexplained by the model.

The first of these information criteria is the -2 Log Likelihood (-2LL). O.K., so what is that? Let me give some background to lead into what it is and what it's used for. The log likelihood itself is computed as part of the maximum likelihood estimation of how well the model fits the data. Again, given the model we propose, we seek to find the values for the parameters (i.e., the value of the intercept, the  $b$ - or  $\beta$ -weights for the factors, etc.) that are most likely given the data; we seek to maximize the chance that those parameters are correct. Without going far into the math of it, we could compute the parameter values by taking the least squares approach for each parameter and then compute the joint probability of getting those parameters (and doing this again and again, tweaking our values to try to make them better and better). However, computing that joint probability can be quite intensive—even for computers. We can lose no precision in our computations but make the math much easier if we instead try to compute the logarithm of those joint probabilities<sup>^</sup>[This is because we could have to multiply the joint probabilities but would only have to add the logs of those probabilities. Adding is less resource-intensive than multiplying.. So, we actually compute the log likelihood instead.

Again, we try to maximize the likelihood of getting those particular parameter values. Similarly, we try to maximize the log of the likelihood (of getting those values). Given how the math works, though, the log likelihood is (nearly always) a negative number. We maximize it by trying to get it out of negative territory—by having it get as close to zero as we can. So, the log likelihood is a negative number, and values closer to zero are better.

However, we multiply the log likelihood by  $-2$  because that transforms it into a value that follows a  $\chi^2$  distribution (so for all of the values in this table, SPSS is right: The information criteria are displayed in smaller-is-better form). This is where the statistics here start to get useful. Remember that the difference between two  $\chi^2$  values also follows a  $\chi^2$  distribution. Therefore, we can take the difference between two  $-2LL$ s and conduct a  $\chi^2$ -test on that difference to see if that difference score is itself significant<sup>23</sup>.

This means that we can take the  $-2LL$  computed for one model, subtract it from the  $-2LL$  computed from another model, and then test if those  $-2LL$ s are significantly different from each other. In other words, if one of those two models fits the data significantly better than the other.

Many researchers, including Singer and Willett, call the  $-2LL$  the *deviance statistic*. This is really the same value (i.e., the difference between a true  $-2LL$  and a deviance statistics can be ignored). Although you will read about deviance statistics more often than “ $-2LL$ s,” I will continue to refer to them as  $-2LL$ s so it’s clearer what part of the SPSS output I’m referring to.

Now, as useful as deviance statistics /  $-2LL$ s are, there are clear limits to when they can be used. The main limitation is that one can only meaningfully compare two models that are computed from the same set of data. Even removing a few cases from a set of data (e.g., by subsetting the data or filter out certain cases), we disqualify tests between them. Yes, this is even a problem if a have missing data that change the sample size between models, so one should take care about that, either removing cases listwise or by imputing values for any missing data.

Singer and Willett also discuss how one of the two models we compare should only contain a subset of the variables of the other model. In other words, one model contains, e.g., gender, race/ethnicity, and IEP status, then we can only compare that to models that either also contain those three variables *plus* other variables (e.g., gender, race/ethnicity, IEP status, *and* executive functioning) or compare it to another model that has a *subset* of those variables (e.g., gender and race/ethnicity). Confusingly, a model that contains a subset of variables from another model is said to be nested in that larger model. So, succinctly, Singer and Willett argue that we can only use  $-2LL$ s to compare models if one is nested within the other.

---

<sup>23</sup>If that  $\chi^2$  difference score is significantly different from zero, to be exact—but that’s a lot of “differences” to digest in one sentence.

<b>Information Criteria<sup>a</sup></b>	
-2 Log Likelihood	28573.945
Akaike's Information Criterion (AIC)	28579.945
Hurvich and Tsai's Criterion (AICC)	28579.953
Bozdogan's Criterion (CAIC)	28600.862
Schwarz's Bayesian Criterion (BIC)	28597.862
The information criteria are displayed in smaller-is-better form.	
a. Dependent Variable: global.exec.comp.sr.	

To compare non-nested models, we use either Akaike's Information Criterion (AIC) or Schwarz's Bayesian Criterion (BIC). We implement either the AIC or BIC the same way we do a deviance statistics: We take the difference between two AICs or between two BICs and see if that difference is significant.

Both the AIC and BIC are based on the  $-2LL$ . The AIC penalizes (makes the absolute value greater) the  $-2LL$  for each factor in a model so that more complex models can't just capitalize on chance. The BIC not only penalizes for the number of factors but also for sample size, so we can't also get better models just by having more data.

You will find either statistic reported in various articles, and either is a fine choice—unless your sample size is so large the BIC becomes necessary. Nonetheless, I tend to use BIC since that is both appropriate even for smaller samples (it simply penalizes them less) and since it is indeed more conservative.

Like the BIC, the Hurvich and Tsai's Criterion (AICC)—more properly written  $AIC_C$ —also attempts to compensate for sample size. The  $AIC_C$ , however, is used for small sample sizes where models may not be able to fully adjust to fit the data.

Bozdogan's Criterion (CAIC) is similar to BIC. It has a stronger penalty for the number of parameters than both the BIC or AIC, but tends to converge on the same values as are computed by the BIC. Given that the BIC is more commonly reported, I think it's fine to stick with the BIC.

Using information criteria to test models is flexible, but we will focus instead here on the more familiar  $F$ - and  $t$ -tests.

### 11.11.5.2 Fixed Effects

The next set of tables are a lot easier to understand. The Type III Test of Fixed Effects presents the  $F$ -score and  $dfs$  testing the significance of the fixed effects in the model. In this unconditional means model, the only term is the intercept, so all we're testing here is whether this value differs significantly from zero.

#### Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	1108.961	27634.189	0.000

a. Dependent Variable: global.exec.comp.sr.

And with an  $F$ -score of 27634.2, I'd say it probably does.

#### Estimates of Fixed Effects<sup>a</sup>

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	152.899427	0.919777	1108.961	166.235	0.000	151.094727	154.704126

a. Dependent Variable: global.exec.comp.sr.

The Estimate of Fixed Effects table reproduces this test (a  $t$ -score of 166.235 is equivalent to an  $F$ -score of 27634.19 since  $166.235^2 = 27634.19$ ). These two tables are redundant because Intercept only has one degree of freedom here. Had we included an interaction term or a term with more than one degree of freedom (e.g., race/ethnicity as a nominal variable with teens coded as "Asian-American," "African-American," "European-American," etc.), then the first Type III Tests table would give us more information about those categories.

This table also presents the Estimate of the global.exec.comp.sr term: 152.90; this is simply the mean score for the entire sample collapsed across waves.

### 11.11.5.3 Covariance Parameters

The covariance parameters are not interesting in this model since they are redundant with the fixed effects.

### 11.11.5.4 Summary

Again, one reason to compute the unconditional mean model are to test if there is an effect overall in the outcome, thus allowing ourselves to further test for effects of time and various predictors. This model also gives us a  $-2LL$  against which we can compare subsequent models, allowing us to see if predictors added to it can improve upon the prediction we would make with just knowing the overall mean for this sample of teens. You may remember we talked about the effects of predictors with just this interpretation of them.

### 11.11.6 Computing the Unconditional Growth Model

Remember that the unconditional growth model tests whether the outcome, here `global.exec.comp.sr`, changes over time, thus warranting further investigations into what other factor may predict there changes. Its goal is thus simple—and so can be our interpretation of it.

To compute this model:

1. Back in the Syntax window, paste in the following syntax:

```
TITLE "Unconditional Growth Model, p. 97".
MIXED global.exec.comp.sr with wave
/PRINT=SOLUTION
/METHOD=ml
/FIXED=wave
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

Note that you can simply paste this in below the syntax for the unconditional mean model.

2. Highlight this syntax and run this syntax. If you're lazy (or just already fighting off the repetitive stress syndrome you'll get from all the typing you'll be doing) instead of highlighting it, you can simply press `Ctrl + A` and then `Ctrl + R` to run all of the commands in the `mlm_syntax.sps` file—another advantage of using syntax and generally learning to use keyboard shortcuts instead of a GUI. (Note you can also simply run a given command when the cursor is placed anywhere within the given command.)

### 11.11.7 Interpreting the Syntax

There are only a few differences in the syntax, but these are profound.

1. The statement is not followed not only by the Outcome (`global.exec.comp.sr`) but then also by the `with wave` statement. The `MIXED` statement now includes an actual model. Models for SPSS syntax are written so that the outcome is predicted with a set of factors. Here, we are only including the wave term, so that alone is added to where the predictors are expected. I need to point out another peculiarity of SPSS's syntax here. In the `MIXED` command, SPSS interprets any factors that follow with `to be random` (i.e., continuous) effects. As we will see later, fixed effects are indicated slightly differently.
2. The `/FIXED` subcommand lets SPSS know which of the predictors in the model given in the `MIXED` command are to be considered fixed factors. Yeah, wave is not a fixed factor—and we would normally not include it here—but we need to have at least one fixed factor, so we add that.
3. In the `/RANDOM` subcommand, we again add wave. It will stay here in subsequent models since it is a random factor—as most all measures of time should be.

### 11.11.8 Interpreting the Results

The Model Dimension table reviews the terms in our model, showing that we now have a wave term added (temporarily as both a fixed and random effect). Note that SPSS indicates that the Number of Levels for the wave term is 1; this will come into play next.

## 11.11.8.1 Information Criteria

The current (unconditional growth) model differs from the previous (unconditional means) model only in that we added a term for wave. This means that the previous model is nested within the current one, and so we can use the  $-2LL$ s to compare the relative fits of these two models to the data.

Information Criteria <sup>a</sup>	
-2 Log Likelihood	28316.912
Akaike's Information Criterion (AIC)	28328.912
Hurvich and Tsai's Criterion (AICC)	28328.941
Bozdogan's Criterion (CAIC)	28370.747
Schwarz's Bayesian Criterion (BIC)	28364.747
The information criteria are displayed in smaller-is-better form.	
a. Dependent Variable: global.exec.comp.sr.	

The  $-2LL$  for the unconditional means model was 28573.945; for the new, unconditional growth model, it is 28316.912.  $28573.945 - 28316.912 = 257.033$ . Therefore, the  $\chi^2$  score for a test between these two models is 257.03. The Number of Levels for the term that was added to this model (compared to the previous, unconditional means model) is 1; this is also the degrees of freedom for this  $\chi^2$  score. So, our tests here is of  $\chi^2 = 257.03$ ,  $df = 1$ . The critical value when  $df = 1$  and a two-tailed  $\alpha = .05$  is  $\chi^2 = 5.02$ . We could justifiably use a one-tailed test since we could argue we're testing if the  $\chi^2$  is greater than zero (and where the critical  $\chi^2 = 3.84$ ), but we'll stick with the two-tailed test.

We can therefore conclude that adding a wave term to the model significant; improves the fit of our model to the data. In other words, we are justified to look further at the ways in which executive functions change here because their changes over time account for a significant portion of the variance in these data.

### 11.11.8.2 Fixed Effects

We do indeed get a similar set of information for the fixed effects tables, where the wave term is significant ( $F_{1, 654.7} = 211.33, p < .001$ ). On average—and across all participants—the global.exec.comp.sr scores change only a little bit: 0.01 points per wave, which is not much for scores that average at 126.6. Nonetheless, these changes are significant and account for a meaningful amount of information herein.

#### Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	393.576	4726.388	0.000
wave	1	654.675	211.335	0.000

a. Dependent Variable: global.exec.comp.sr.

#### Estimates of Fixed Effects<sup>a</sup>

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	126.576791	1.841151	393.576	68.749	0.000	122.957070	130.196511
wave	0.011286	0.000776	654.675	14.537	0.000	0.009762	0.012811

### 11.11.8.3 Covariance Parameters

The covariance parameters table now includes some more information. The Residual in this table reports statistics related to the variance within each participant. The Estimate for the residual variance is quite high, indicating that a lot of variance is still unexplained. This is equivalent to Singer and Willett's level 1 model [residual](#). Frankly, this Residual Estimate isn't commonly reported and is only marginally informative for one's in-house analyses. It does convey the amount of unexplained variance left after the given model, but so do the information criteria, and those are presented in a format that is useful across models conducted on the same data in ways that are more amenable to further analyses.

## Estimates of Covariance Parameters<sup>a</sup>

Parameter		Estimate	Std. Error
Residual		646.466673	24.159758
Intercept + wave [subject = id]	UN (1,1)	438.156300	140.608569
	UN (2,1)	-0.106359	0.058766
	UN (2,2)	0.000104	2.651810E-05

a. Dependent Variable: global.exec.comp.sr.

The three rows within the Intercept + wave [subject = id] section provide parameter estimates (and standard errors for those estimates) for terms that are between participants. These are terms for random effects that are part of Singer and Willett's [level 2 model](#). The three terms reported herein are the variance of the intercept (UN (1,1)), covariance between the intercept and wave terms (UN (2,1)), and variance of the wave term (UN (2,2)). These terms are useful to investigate how much of the residual variance remains in the various areas, but—again—further understanding of the relationships are probably better studied through careful analyses of the predictors and perhaps graphs of the residuals.

One thing to point out from these level-2 variances is that the covariance between the intercepts and changes over waves is “borderline” significant ( $p = .059$ ). This suggests that there may not be a significant relationship between adolescents' initial levels of executive functioning and subsequent changes in it.

We can see from these residuals that, although intercept remains highly predictive, there is still much we don't understand about what affects that values. This does nonetheless reflect a large body of research indicating that much of one's executive functioning is determined during childhood.

### 11.11.8.4 Summary

The unconditional change model is the second of two models that simply help establish the foundation upon which other models can be built and against which they can be compared. We do see through these two models that there are non-negligible inter-individual differences and that executive functioning does appreciably change over these administrations.

## 11.12 Univariate MLMs

We will next review a series—or “taxonomy”—of models each that contains one predictor (in addition to those in the unconditional models). Even with this rather limited set of variables, there are many comparisons we could make. However, we'll keep it simple and look at the relative contributions of only three predictors on global.exec.comp.sr: gender, iep.status, and economic.distress. We will presume that the variable of especial interest here is economic.distress and whether it makes a unique contribution beyond that already made by gender and iep.status.

To investigate these three predictors, we will first look at the relationship of each one with `global.exec.comp.sr` alone, i.e., without the other two predictors in the model. We will also look at the relationship between each predictor and the adolescents' initial levels of executive functioning as well as the relationship with subsequent changes in executive functioning. Each of the univariate models therefore looks at the contribution of the given predictor without consideration of that predictor's relationship (correlation) with any of the other predictors.

These univariate models are also themselves “prefatory” ones that simply help lay the groundwork for the final few models that we would likely actually report in an manuscript. We are, of course, looking here at a few pieces of what is likely a very tangled web of influences on the lives and development of these teens. In order to understand any of it we are well-advised to first try to look at the pieces in relative isolation before considering how they may interact. We will do that now.

### 11.12.1 Gender

#### 11.12.1.1 Main Effect

We will first look at whether boys' and girls' initial `global.exec.comp.sr` scores differ.

#### Via SPSS's GUI

We can—and will—do this using the syntax, but the model is complex enough to serve as an example for how to do this using the GUI:

Please note that we can run a similar model via the GUI:

1. Click on Analyze > Mixed Models > ``Linear...
2. Add `id` to the Subjects: field, `wave` to the Repeated: field, and change the Repeated Covariance Type: to Unstructured
3. Click Continue at the bottom
4. In the next dialogue, add `global.exec.comp.sr` to the Dependent Variable: field, `gender` to the Factor(s): field, and `wave` to the Covariates: field
  1. Under Fixed..., change the middle button's choice to Main Effects, add `gender` to the Model:, and make sure Include intercept is selected (and that the Sum of squares is Type III). Adding the intercept here lets each participant and each gender to have their own beginning value.
  2. Under Random..., leave/make the middle button selection as after selecting Factorial and add `wave` to the Model: field.
  3. Under Estimation..., chose Maximum Likelihood (ML) under Method and leave all other values at their defaults.
  4. Under Statistics..., in Model Statistics, select Parameter estimates for fixed effects and Tests for covariance parameters.
  5. Under EM Means..., add `gender` to the Display means for: field.
  6. Finally, click OK to run.

### Via Syntax

We do this by adding only a gender term to the model. To do this, run the following syntax:

```
TITLE "Gender Main Effect".
MIXED global.exec.comp.sr with wave by gender
/PRINT=SOLUTION TESTCOV
/EMMEANS=TABLES(gender)
/METHOD=ml
/FIXED=gender
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

### Interpreting the Syntax

The lines that differ are MIXED and /FIXED. We've also added a /EMMEANS=TABLES(gender) subcommand.

The MIXED command again contains the Outcome, but now followed by both a with and a by argument<sup>24</sup>. The with argument denotes the factors that should be considered as random effects (here, wave) while the by argument denotes factors that should be considered as fixed effects, like gender. Sure, gender could be seen as not measuring all possible levels of that domain, but we'll treat it as fixed since that's how it's measured by the school.

In the /FIXED subcommand, we are also just adding a "main effect" of gender. This will become clearer in the next model, but for now know that we are only looking at the effect of gender on the initial (intercept) Outcome scores.

The /RANDOM subcommand has remained the same. We are again noting that wave is a random factor and that it is to be considered nested within id.

The /EMMEANS=TABLES(gender) subcommand requests that SPSS print a table of the estimated marginal means for the predictor(s) along with their standard errors and 95% confidence intervals.

The estimated marginal means are simply the means for the various levels of the given predictors that are predicted by the model (i.e., *not* their actual means), after partialing out the effects of any other terms. We only have one predictor in the current, so we are only partialling out the intercept here.

### Interpreting the Results

At last we are looking at theoretically-interesting results. The point of the current analysis is to investigate the relationship between the adolescents' gender and their executive functioning.

We can look at it here in two ways: the significance of gender's model term and at the change in model fit when gender is added to the previous model. We will focus on the former, looking at the significance of the gender term itself.

The Model Dimension table summarizes the model variables, covariance structure, degrees of freedom, and which variables are nested within which:

---

<sup>24</sup>This is one of the main ways that my syntax differs from that posted in the [companion website](#) to Singer and Willett's book: They don't separate out effects into by statements.

Model Dimension <sup>a</sup>		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	gender	2		1	
Random Effects	Intercept + wave <sup>b</sup>	2	Unstructured	3	id
Residual				1	
Total		5		6	

a. Dependent Variable: global.exec.comp.sr.

b. As of version 11.5, the syntax rules for the RANDOM subcommand have changed. Your command syntax may yield results that differ from those produced by prior versions. If you are using version 11 syntax, please consult the current syntax reference guide for more information.

In the Number of Parameters column, we can see in that table that the nominal gender variable contains two levels (male and female). Note, though, that in the Number of Parameters column, there is only one “parameter”; this is the number of degrees of freedom used to add this term to the model. The number of parameters will always be one less than the number of levels for that term.

Note, too, that both the Intercept and gender terms are listed under Fixed Effects. Again, SPSS considers all fixed effects to be between participants (in Singer and Willett’s level 2 model). wave remains under the Random Effects; SPSS considers all random effects to be within participants.

### Fixed Effects

Intentionally, we have only one fixed effect added to the model, allowing us to investigate that effect in isolation. The effect we’re testing is whether girls and boys began this study (in sixth grade) with different levels of executive functioning.

### Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	903.779	23727.448	0.000
gender	1	903.779	3.049	0.081

a. Dependent Variable: global.exec.comp.sr.

The Type III Tests of Fixed Effects table also tests the effect of gender. This test is tantamount that one would compute in an ANOVA—although here we are using maximum likelihood estimation<sup>25</sup>—so this *F*-score tests for mean difference between the gender groups. This *F*-score is *not* significant ( $F_{1,903.8} = 3.05$ ,  $p = .081$ ). There is insufficient evidence that boys and girls began this study in sixth grade self-reporting different levels of executive functioning.

<sup>25</sup>The ordinary least squares estimation one uses in an ANOVA arrives at the same outcome as maximum likelihood estimation—when the assumptions of the ANOVA are met (viz., when residuals are truly normally distributed).

**Estimates of Fixed Effects<sup>a</sup>**

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	154.010584	1.413383	920.208	108.966	0.000	151.236756	156.784412
[gender=0]	-3.452399	1.977242	903.779	-1.746	0.081	-7.332920	0.428121
[gender=1]	0 <sup>b</sup>	0					

a. Dependent Variable: global.exec.comp.sr.

b. This parameter is set to zero because it is redundant.

The Estimates of Fixed Effects table that comes next reflects the same results as the estimated marginal means because the predictor has only two levels. Note that there is an oddity to this table: The gender term is divided into two rows, with values given for [gender=0] but not [gender=1]<sup>26</sup>. SPSS is assuming that the highest category for this predictor (when gender = 1) is the reference group—the level against which the other category is compared. So, the test here is whether those participants whose gender = 0 (boys) have significantly different mean executive functioning scores than the reference group (girls).

The Estimates of Fixed Effects table thus indicates that boys' scores are 3.45 points lower than girls' scores. This value of -3.45 is also the  $\beta$ -weight for the gender term (sigh, when girls are defined as the reference group; were I to report this, I would reverse the sign to meet readers' expectations).

One more thing to say about comparisons like the levels of gender in the Estimates of Fixed Effects table: SPSS always assumes the highest level is the reference group. Or at least that I don't know how to change it away from that. This is odd and inconvenient since dummy variables are designed to have the group that's zeros be the reference group.

The Estimates of Covariance Parameters table still presents the within-participant residual variance in the first row and variances for the intercept, intercept  $\times$  slope interaction, and slope per se, respectively. We needn't reproduce that table here not consider it further in this context.

**Gender<sup>a</sup>**

Gender	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
0	150.558	1.383	885.804	147.844	153.272
1	154.011	1.413	920.208	151.237	156.784

a. Dependent Variable: global.exec.comp.sr.

A new table has appeared after that Covariance table, here called simply Gender. These are the estimated marginal means requested by the /EMMEANS=TABLES(gender) subcommand that we added to this analysis. Again, this provides the initial global.exec.comp.sr scores the model estimates boys and girls had at the time defined as crossing the intercept. We therefore estimate that boys'

<sup>26</sup>It is my own convention to always set dichotomized gender so that female = 1 (and male = 0) whenever it is dichotomized as such, but always following the same routine makes it easier to remember and interpret. This also seems to me to give a bit more prominence to women, which is never a bad thing. Being dichotomized, though, we know nothing about the other vast and varied facets of sexual identity.

initial scores were 150.558 and girls' were 154.011. Note that  $150.558 - 154.011 = -3.45$ , which is the Estimates value for boys ([gender=0]) in the Estimates of Fixed Effects table.

## Summary

A multilevel model of change did not find that gender significantly predicted under-served adolescent students' initial levels of self-reported executive functioning ( $F_{(1, 907.8)} = 3.05$ ,  $p = .081$ ) when that term was included with no other predictors except terms for the intercept and time. But we are not done with gender yet. We did not find that girls and boys started middle school with different levels of executive functioning, but this does not mean that they will not change as time goes by. Let us now look into that.

### 11.12.1.2 Gender × Time Interaction

A strategy for investigating factors in this fashion includes one piece of information at a time to see what that adds to our understanding: We proceed in careful, precise steps to ensure accurate understanding before building those pieces together into a larger picture. Therefore, we will look now only at how gender may affect changes in executive functioning over time—independent of any effect gender (could have) had on the initial levels of executive functioning.

To do this, we simply add only a gender × wave interaction term to the “null” comparison model that contains only intercept and wave terms with the following syntax:

```
TITLE "Gender Interaction".
MIXED global.exec.comp.sr with wave by gender
/PRINT=SOLUTION TESTCOV
/EMMEANS=TABLES(gender*wave)
/METHOD=ml
/FIXED=gender*wave
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

## Interpreting the Syntax

The only difference between this model and the one we analyzed in the Gender Main Effect section is that the `/FIXED=gender*wave` line where we have a gender × wave interaction term added instead of a gender main effect term.

## Interpreting the Results

### Information Criteria

We will not review the information criteria for this model. Neither it nor the previous model containing the gender main effect are nested within each other. We could use AIC (or BIC) here to compare whether gender predicting intercept accounts for more information in the data than gender predicting changes in executive functioning, but this is not of great interest here. In addition, the test of the term itself suffices to study its effects.

**Fixed Effects****Type III Tests of Fixed Effects<sup>a</sup>**

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	366.638	4305.118	0.000
gender * wave	2	649.561	104.929	0.000

a. Dependent Variable: global.exec.comp.sr.

The gender  $\times$  wave interaction is significant ( $F_{2, 649.6} = 104.93, p < .001$ ). Note that this is an interaction between a variable nested within participant (wave) and an other variable that is between participants (gender), but this still appears in the Fixed Effects tables. It is interpreted straight forwardly that the slopes of the executive functioning scores is significantly different for the boys and girls.

Estimates of Fixed Effects <sup>a</sup>						
Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound
Intercept	125.126569	1.907028	366.638	65.613	0.000	121.376484
[gender=0] * wave	0.011490	0.000940	692.214	12.225	0.000	0.009644
[gender=1] * wave	0.013031	0.000950	732.006	13.715	0.000	0.011165

The Estimates of Fixed Effects table gives us insight into the nature of this interaction: On average, boys' executive functioning scores increase 0.011 points per wave (here, that's essentially per academic year) while girls' scores increase slightly more, 0.013. Both of these increases—small as they are—are significant here. The standard errors help explain why: There is not much variance in the rates of change, so even small effects are detectable.

Note that the Estimates of Fixed Effects table presents results for analyses of the individual levels of the gender  $\times$  wave interaction. We are essentially conducting post hoc comparisons of the levels to see if both are significant. If the  $F$ -test given in the Type III Tests of Fixed Effects table was not significant, then these  $t$ -tests of the levels are not warranted. SPSS would report them anyway, though, so only interpret the level tests if the  $F$ -test is first significant.

**Covariance Parameters****Estimates of Covariance Parameters<sup>a</sup>**

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		647.204912	24.453276	26.467	0.000	601.008957	696.951673
Intercept + wave [subject = id]	UN (1,1)	459.312693	144.317828	3.183	0.001	248.117630	850.274727
	UN (2,1)	-0.132000	0.062326	-2.118	0.034	-0.254157	-0.009844
	UN (2,2)	0.000125	2.929820E-05	4.280	0.000	7.932629E-05	0.000198

a. Dependent Variable: global.exec.comp.sr.

An additional piece of information become interesting in the Estimates of Covariance Parameters table. Remember that the UN(2,1) row presents the covariance between the intercept and time terms. Using a Wald test, this relationship is significant ( $W = 4.28$ ,  $SE < 0.001$ ,  $p < .001$ ) and positive ( $b = 0.00012$ ), if rather small. This suggests that teens with larger executive functioning scores tended to have (slightly) more positive slopes than teens with smaller scores.

## Summary

The rather small variance in the rates of change in executive functioning allowed us to investigate their relationships with gender in some detail. We found that overall adolescents with grater executive functioning scores tended to have score that got even greater throughout their middle and high school grades while adolescents with lower scores tended to have scores that further decreased: To a small but significant degree, small initial differences in executive functioning tended to grow in difference.

We also found that girls' executive functioning scores tended to become slightly greater over time. These executive functioning scores are coded so that lower scores denote stronger executive functioning, so—surprisingly—girls tended to show a greater reduction in their self-reported executive functioning relative to boys. Findings that disconfirm our expectations tend to garner ore attention (or, well, they should), so it may be worth, e.g., looking at comparable executive functioning scores reported by teachers about these same students; perhaps the effect here is not replicated in how some others view these teens' executive-functioning-related behaviors. Perhaps instead the changes in self-reported scores are more related to changes in, e.g., one's self confidence.

## 11.12.2 Special Education

We next next review a similar pair of models to investigate the effects of IEP status on both initial executive functioning levels and subsequent changes thereof. Just as a reminder, an IEP is an “individualized education program” (or “plan”) designed to help address students with special needs that have been diagnosed to affect their academic performance.

### 11.12.2.1 Main Effect

One advantage of using syntax is that it is sometimes easy to conduct a bevy of analyses while making only a few small changes much more quickly than one could do through a GUI. Here, we need only change the title (for future reference) and change instance of the word “gender” to “iep.status”.

Or, simply paste the following syntax into the Syntax Editor:

```
TITLE "IEP Status Main Effect".
MIXED global.exec.comp.sr with wave by iep.status
/PRINT=SOLUTION TESTCOV
/EMMEANS=TABLES(iep.status)
/METHOD=ml
/FIXED=iep.status
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

### Interpreting the Syntax

This syntax will conduct the same “main effect” analysis testing whether students with and without IEPs differ significantly in their initial, sixth-grade levels of self-reported executive functioning. The only changes in it (outside of the title) are changing gender to `iep.status`. a change one could make by searching and replacing that phrase, but then also reviewing the code to ensure doing it automatically didn’t have any unintended consequences elsewhere in the code.

### Interpreting the Results

Since we have already detailed the particulars of the results when investigating gender, we will be more brief in our coverage of `iep.status`. Hopefully this will help reinforce the main points and encourage practice remembering.

### Information Criteria

We can test the change in fit of the entire model through the information criteria. We could use the  $-2LL$  (deviance) statistics to compare this to the unconditional means or growth model since both of those models are nested within the current: They contain only subset of the terms in the current model.

The current model has one more term than the unconditional growth model—`iep.status`—which the Model Dimension table tells me adds only one to the Number of Parameters. Therefore, I could test for a change in model fit by subtracting the  $-2LL$  for the current model from the  $-2LL$  for the unconditional growth model and testing that difference against a critical  $\chi^2$  value with one degree of freedom.

The current model has *two* more terms than the unconditional means model: the `iep.status` term and the wave term that was added in the unconditional growth model. The current model thus has two more degrees of freedom than the unconditional means model<sup>27</sup>, so the difference in  $-2LL$ s would be compared against the critical  $\chi^2$  value for 2 *dfs*.

### Fixed Effects

#### Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	883.063	21025.459	0.000
iep.status	1	883.063	13.357	0.000

a. Dependent Variable: `global.exec.comp.sr`.

<sup>27</sup>It is not always the case that a new term has only one degree of freedom. Interaction terms rarely will, and the ethnicity nominal variable has several. Check the Model Dimensions table to see how many degrees of freedom—Number of Parameters—each term has.

Estimates of Fixed Effects <sup>a</sup>							
Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	158.297275	1.746063	889.950	90.660	0.000	154.870394	161.724155
[iep.status=0]	-7.783401	2.129709	883.063	-3.655	0.000	-11.963282	-3.603519
[iep.status=1]	0 <sup>b</sup>	0					

a. Dependent Variable: global.exec.comp.sr.

The Type III Tests of Fixed Effects and the Estimates of Fixed Effects tables indicate that the iep.status term is significant.

Since iep.status scored as a dummy variable, the levels for it given in Estimates of Fixed Effects table are interpreted somewhat unconventionally: there are zeros in the row for [iep.status=1] (denoting that the student has an IEP). We can nonetheless use this table to interpret the magnitude of the effect of having an IEP since those without one ([iep.status=0]) begin the study with an average of 7.8 points less on the scale (the BRIEF-SR GEC) than those with an IEP; since lower scores denote stronger executive functioning, this suggests that those without an IEP already have stronger executives functions than those with an IEP.

### Covariance Parameters

Estimates of Covariance Parameters <sup>a</sup>							
Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		625.639615	23.909614	26.167	0.000	580.489663	674.301289
Intercept + wave [subject = id]	UN (1,1)	1579.112037	242.497420	6.512	0.000	1168.685341	2133.675112
	UN (2,1)	-0.557151	0.096787	-5.756	0.000	-0.746851	-0.367451
	UN (2,2)	0.000286	4.135712E-05	6.917	0.000	0.000215	0.000380

a. Dependent Variable: global.exec.comp.sr.

The Estimates of Covariance Parameters table presents the residuals at both the between-participant level (level 2) as Residual and the variances and covariances of the within-participant (level 1) terms below that. This table shows again that there is significant variance yet unexplained and that the intercepts covary significantly with the changes—note, though, that this is *not* changes in outcomes related to IEP status since we have not included an iep.status × wave interaction term to the model; this is simply whether intercept and slope are related in the overall set of data.

IEP Status <sup>a</sup>					
IEP Status	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
0	150.514	1.219	868.648	148.121	152.907
1	158.297	1.746	889.950	154.870	161.724

a. Dependent Variable: global.exec.comp.sr.

The Estimated Marginal Means table—entitled by the Label for the variable reported—indicates that the model estimates that the mean `global.exec.comp.sr` score for students without IEPs is 150.5, which is 7.8 points lower than the mean score estimated for those with IEPs.

### Summary

We found that IEP status is significantly related to initial levels of executive functioning. Indeed, it appears rather predictive of this initial level given the magnitude of the effect.

#### 11.12.2.2 IEP Status $\times$ Time Interaction

We now look at the effect of IEP status on changes in executive functioning over time. Looking at it in a separate model like this helps compare differences in model fit. It also lets us compare how much the effect of IEP on intercept is related to the effect of IEP status in subsequent changes in the outcome because we look at both effects in isolation and then can look at them together in the same model.

Once again, we simply change `gender` to `iep.status` in the model containing the interaction term:

```
TITLE "IEP Status Interaction".
MIXED global.exec.comp.sr with wave by iep.status
/PRINT=SOLUTION TESTCOV
/EMMEANS=TABLES(iep.status)
/METHOD=ml
/FIXED=iep.status iep.status*wave
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

### Interpreting the Results

#### Information Criteria

Note in the Model Dimension table that the `iep.status  $\times$  wave` interaction term has 2 degrees of freedom—not 1. An additional degree of freedom is used to account for interactions, but only one for the first interaction; if we had added two interactions *with the same within-participant term*, we wouldn't need to continue to add more than one degree of freedom per interaction. In other words, if we had included both an `iep.status  $\times$  wave` and a `gender  $\times$  wave` interaction, we would only need to add the additional degree of freedom (for the interaction with wave) once—not twice. This is confusing, can better seen by an example, which I will when we compare more complex models, below.

#### Fixed Effects

### Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	333.087	3864.332	0.000
iep.status * wave	2	637.015	91.466	0.000

a. Dependent Variable: global.exec.comp.sr.

Estimates of Fixed Effects <sup>a</sup>						
Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound
Intercept	126.420921	2.033676	333.087	62.164	0.000	122.420454
[iep.status=0] * wave	0.010885	0.000935	590.253	11.637	0.000	0.009048
[iep.status=1] * wave	0.013278	0.001050	738.966	12.649	0.000	0.011217

IEP status significantly predicts changes in executive functioning over the middle and high school years among these adolescents. The relationship is rather strong even though the effect is still not large: The global.exec.comp.sr scores for those with IEPs move toward levels indicating worse executive functioning at the mean rate of about 0.013 points per year, compared to those without IEPs whose scores move in the same direction a bit more slowly (at the rate of about 0.011 points per year).

#### Covariance Parameters

The pattern of significances among the random effect terms when  $iep.status \times wave$  is added to the model reflects that found when  $iep.status$  was added. The values differ, e.g., for the intercept  $\times$  change term (UN(2,1)) because we are estimating the effect with somewhat different information.

#### Summary

Not only do those students with IEPs already come into the range of grades measured here with significantly worse self-reported executive functioning, but they also tend to show significantly greater depreciation in those faculties throughout their secondary grades.

We chose to investigate executive functioning because we anticipated nurturing its development could help those who are struggling to overcome challenges both within and without. Seeing executive functioning weaken most among those who may benefit most from it seems alarming.

### 11.12.3 Economic Distress

And yet among all of the challenges faced by these teens, their economic ones may be most pervasive. Like race/ethnicity, the effect of poverty is often misattributed. Having relatively little money is among the least of the challenges faced by the poor; more salient is living in less safe, more stressful, and more dangerous conditions; limited access to healthy food and lifestyle

choices; fewer opportunities for success and many voices taking their failure for granted; and several other factors that could affect not only the development of their executive functioning, but also whether they have disabilities warranting IEPs and whether they are correctly diagnosed with those needs. All of this may also be filtered through the increasing different experiences of boys and girls in these environments.

In short, economic.distress may affect initial levels of executive functioning, subsequent rates of development. It may also interact with one's special needs—and this all may be filtered through the lens of being a boy or girl.

### 11.12.3.1 Main Effect

We are presuming here that our main interest is indeed on economic distress and whether any relationship it has with executive functioning is moderated by IEP status and gender—or the extent to which its effects are independent of those other factors.

We begin this phase of our analysis by first looking at the relationships of economic distress alone, without consideration of the effects of IEP status or gender. And first, the relationship with initial levels:

```
TITLE "Economic Distress Main Effect".
MIXED global.exec.comp.sr with wave by economic.distress
/PRINT=SOLUTION TESTCOV
/EMMEANS=TABLES(economic.distress)
/METHOD=ml
/FIXED=economic.distress
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

### Interpreting the Results

#### Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	961.562	23391.502	0.000
economic.distress	1	961.562	3.357	0.067

a. Dependent Variable: global.exec.comp.sr.

#### Estimates of Fixed Effects<sup>a</sup>

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound
Intercept	150.352543	1.274216	826.473	117.996	0.000	147.851462
[economic.distress=0]	3.646027	1.989967	961.562	1.832	0.067	-0.259152
[economic.distress=1]	0 <sup>b</sup>	0				

a. Dependent Variable: global.exec.comp.sr.

Unexpectedly (after my harangue at least), economic.distress does not have a significant effect on the initial levels of executive functioning.

### 11.12.3.2 IEP Status × Time Interaction

We now investigate the effects of economic distress on subsequent changes in executive functioning with the following syntax:

```
TITLE "Economic Distress Interaction".
MIXED global.exec.comp.sr with wave by economic.distress
/PRINT=SOLUTION TESTCOV
/EMMEANS=TABLES(economic.distress)
/METHOD=ml
/FIXED=economic.distress*wave
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

### Interpreting the Results

#### Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	378.180	4416.693	0.000
economic.distress * wave	2	673.944	103.580	0.000

a. Dependent Variable: global.exec.comp.sr.

#### Estimates of Fixed Effects<sup>a</sup>

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
Intercept	125.255276	1.884723	378.180	66.458	0.000	121.549428	128.961124
[economic.distress=0]* wave	0.012766	0.000968	803.997	13.189	0.000	0.010866	0.014666
[economic.distress=1]* wave	0.011628	0.000917	660.991	12.676	0.000	0.009827	0.013429

a. Dependent Variable: global.exec.comp.sr.

Skipping to the variable-level tests, we see that—taken in isolation—economic distress has a clearly significant effect on the development of executive functioning throughout middle and high school. The difference in slopes for those above and below the criterion for being considered “economically distressed” show only slightly different trajectories, but the standard errors again indicate why this is significant.

### 11.12.4 Summary of Univariate Model Analyses

The adolescents’ initial levels of executive functioning were significantly predicted by whether the teen had an IEP, but not by the teen’s gender nor whether they were experiencing economic

distress. All of the predictors were significantly related to changes in executive functioning over time.

These results, however, were found when looking at each relationship separately. It is entirely plausible that they will interact. For example, the effect of economic distress may be intermixed with the effect of having an IEP, and gender may moderate either effect.

Even among this limited set of variables, there are many further analyses we can conduct to better understand what is happening in these teens' lives. We will look only at one specific scenario, though: How the effects that we found significant in isolation interact when included together.

We will look at the significant relationships in two steps. First, we will include the gender and IEP status effects together in a base model. We will then add in economic distress's interaction to this model to see whether this adds new information to our understanding.

### 11.13 Multivariate MLM

We will now demonstrate using multiple predictors, thus creating a multivariate MLM.

Since we are now using more than one predictor, there is the chance of multicollinearity affecting the value (and stability) of the model parameters—especially for those predictors that are highly correlated with each other. Although there are analyses that can detect multicollinearity in multi-level (hierarchical) models (q.v., Yu et al., 2015), reviewing them is outside of the pale of this course. Note that in general, multicollinearity is usually less of a problem than it's sometimes conceived to be, and that it matters in MLMs more when it is between nested groups than within one.

#### 11.13.1 Base Model

We create our base, comparison model with:

```
TITLE "Gender and IEP Status".
MIXED global.exec.comp.sr with wave by iep.status gender
/PRINT=SOLUTION TESTCOV
/METHOD=ml
/FIXED=iep.status iep.status*wave gender*wave
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

This includes main effect and interaction terms for iep.status and the interaction term for gender.

##### 11.13.1.1 Interpreting the Results

#### Information Criteria

		Model Dimension <sup>a</sup>			
		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	iep.status	2		1	
	iep.status * wave	2		2	
	gender * wave	2		1	
Random Effects	Intercept + wave <sup>b</sup>	2	Unstructured	3	id
Residual				1	
Total		9		9	

a. Dependent Variable: global.exec.comp.sr.

b. As of version 11.5, the syntax rules for the RANDOM subcommand have changed. Your command syntax may yield results that differ from those produced by prior versions. If you are using version 11 syntax, please consult the current syntax reference guide for more information.

We will return to using the information criteria (viz., the  $-2LL$ —or deviance—statistic) in addition to reviewing the tests of the individual parameters. First, note that this model is using 9 *dfs* to estimate the values for the parameters.

We can see from this Model Dimension table that a degree of freedom (a parameter) was only added once—here listed in the `iep.status * wave` row, but it would have been listed in the `gender * wave` row if we had added `gender*wave` first in the syntax. This is because we only need to add that parameter once for all interactions with that within-participant term (`wave`). When we add the `economic.distress*wave` term, below, we will thus only be adding one additional parameter—not two.

## Information Criteria<sup>a</sup>

-2 Log Likelihood	25594.418
Akaike's Information Criterion (AIC)	25612.418
Hurvich and Tsai's Criterion (AICC)	25612.487
Bozdogan's Criterion (CAIC)	25674.270
Schwarz's Bayesian Criterion (BIC)	25665.270

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: global.exec.comp.sr.

Also note that the  $-2LL$  for this base model is 25594.418.

### Fixed Effects

## Type III Tests of Fixed Effects<sup>a</sup>

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	303.270	3572.759	0.000
iep.status	1	303.311	9.953	0.002
iep.status * wave	1	489.389	1.763	0.185
gender * wave	1	725.291	5.542	0.019

a. Dependent Variable: global.exec.comp.sr.

**Estimates of Fixed Effects<sup>a</sup>**

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	135.186759	3.496526	285.119	38.663	0.000	128.304479	142.069039
[iep.status=0]	-13.555118	4.296566	303.311	-3.155	0.002	-22.009969	-5.100266
[iep.status=1]	0 <sup>b</sup>	0					
[iep.status=0] * wave	0.013726	0.001158	626.153	11.849	0.000	0.011451	0.016001
[iep.status=1] * wave	0.011241	0.001596	530.995	7.042	0.000	0.008105	0.014377
[gender=0] * wave	-0.001944	0.000826	725.291	-2.354	0.019	-0.003565	-0.000323
[gender=1] * wave	0 <sup>b</sup>	0					

a. Dependent Variable: global.exec.comp.sr.

b. This parameter is set to zero because it is redundant.

When all three terms are added to the model, the `iep.status * wave` interaction is no longer significant. Both of those with and without IEPs show changes in executive functioning, but since the omnibus *F*-test did not find that term to be significant, those changes over time are not significantly different.

### Covariance Parameters

We are not analyzing the covariance terms here. Since the model is more complex, the weights among the residuals and within-participant variances and covariances are harder to interpret (without direct comparisons to other models to disentangle these effects).

### Summary

This is the base model, against which we wish to compare the effects of economic distress. Therefore, the significances of the terms are not of primary interest here. Nonetheless, it is interesting to see that changes over time due to IEP status are no longer significant when we also account for changes due to gender.

### 11.13.2 Final Model

To this base model, we now add one addition term: the `economic.distress * wave` interaction. We are therefore testing here whether the economic distress contributes significantly to our understanding of adolescent development of executive functioning beyond that already made by the other terms. If economic distress is significant, then the effect it has on development is at least partly due to factors independent of gender and IEP status. If economic distress is not significant, then the relationship we found earlier between it and executive functioning development may be sufficiently accountable by gender and IEP status.

The syntax for this final model is:

```
TITLE "Adding Eco Dis Interaction to Model w/ Gender & IEP Status".
MIXED global.exec.comp.sr with wave by economic.distress iep.status gender
/PRINT=SOLUTION
/METHOD=ml
```

```
/FIXED=economic.distress*wave iep.status iep.status*wave gender*wave
/RANDOM intercept wave | SUBJECT(id) COVTYPE(un).
```

In this syntax, we have added `economic.distress` to the `MIXED` command and the `economic.distress * wave` interaction to the list of `/FIXED` parameters.

### 11.13.2.1 Interpreting the Results

#### Information Criteria

		Model Dimension <sup>a</sup>		
		Number of Levels	Covariance Structure	Number of Parameters
Fixed Effects	Intercept	1		1
	economic.distress * wave	2		2
	iep.status	2		1
	iep.status * wave	2		1
	gender * wave	2		1
Random Effects	Intercept + wave <sup>b</sup>	2	Unstructured	3
Residual				1
Total		11		10
a. Dependent Variable: global.exec.comp.sr.				

Note that we only added one additional parameter: The base model (without the `economic.distress * wave` interaction) had 9 Total parameters whereas the current model now has 10.

We can see in the Information Criteria table just below that the  $-2LL$  for this extended model is 25592.703. The  $2LL$  for the base model was 25594.418; the difference between these two models is  $25594.418 - 25592.703 = 1.75$ . The critical  $\chi^2$  for 1 *df* is 5.02 for a two-tailed test, or 3.84 for a one-tailed test; in either case, the economic distress term did not make a significant contribution to the fit of the model to the data.

## Information Criteria<sup>a</sup>

-2 Log Likelihood	25592.703
Akaike's Information Criterion (AIC)	25612.703
Hurvich and Tsai's Criterion (AICC)	25612.787
Bozdogan's Criterion (CAIC)	25681.427
Schwarz's Bayesian Criterion (BIC)	25671.427

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: global.exec.comp.sr.

### Fixed Effects

This non-significance is reflected in the variable-level tests:

Type III Tests of Fixed Effects <sup>a</sup>				
Source	Numerator df	Denominator df	F	Sig.
Intercept	1	305.503	3570.347	0.000
economic.distress * wave	1	717.807	1.724	0.190
iep.status	1	304.750	9.701	0.002
iep.status * wave	1	490.404	1.691	0.194
gender * wave	1	724.980	5.588	0.018

a. Dependent Variable: global.exec.comp.sr.

The omnibus  $F$ -test did not find the `economic.distress * wave` term to be significant. Therefore, the additional tests of the levels of the `economic.distress * wave` interaction are not warranted (even though they are still given):

Estimates of Fixed Effects <sup>a</sup>							
Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
Intercept	135.208848	3.498685	286.777	38.646	0.000	128.322488	142.095207
[ <code>economic.distress=0</code> ] * wave	0.011876	0.001669	599.488	7.114	0.000	0.008597	0.015155
[ <code>economic.distress=1</code> ] * wave	0.010784	0.001635	553.719	6.596	0.000	0.007573	0.013996
[ <code>iep.status=0</code> ]	-13.396291	4.301169	304.750	-3.115	0.002	-21.860041	-4.932541
[ <code>iep.status=1</code> ]	0 <sup>b</sup>	0					
[ <code>iep.status=0</code> ] * wave	0.002436	0.001873	490.404	1.300	0.194	-0.001245	0.006117
[ <code>iep.status=1</code> ] * wave	0 <sup>b</sup>	0					
[ <code>gender=0</code> ] * wave	-0.001950	0.000825	724.980	-2.364	0.018	-0.003570	-0.000331
[ <code>gender=1</code> ] * wave	0 <sup>b</sup>	0					

a. Dependent Variable: `global.exec.comp.sr`.

b. This parameter is set to zero because it is redundant.

## Summary

We know from our univariate analysis that economic distress—when taken alone—significantly predicts changes in executive functioning. However, we now see from this analysis that the effect of economic distress here is arguably explainable through the effects of IEP status and/or gender.

Of course, this oughtn't be the final word on the issue. We do not know yet if the covariance between economic distress and changes in executive functioning overlaps that of IEP status, gender, or both. We could look at this by next taking out those terms systematically. We could also, e.g., add an `economic.distress * gender * wave` interaction to test the effects of more complex relationships.

Wholly beyond the pale of the current chapter would be to see if perhaps `iep.status` (or maybe even gender) mediate the relationship between `economic.distress` and `wave`. That will require another course to discuss—perhaps one on measurement and factor analysis....

Monsalves et al. (2020) provides good advice on writing up models like these. A couple other guides are given in the Resources section of Chapter 3: Writing Results Sections.

## 11.14 Additional Resources & Topics

### 11.14.1 Some Other Ways to Analyze Longitudinal Data

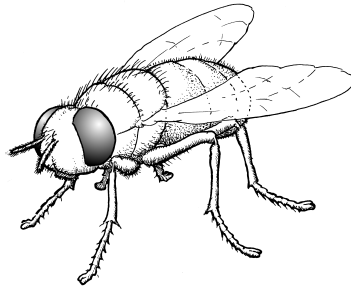
#### 11.14.1.1 Statistical Process Control

Statistical process control is a more graphical approach to longitudinal analyses. It focuses on visually scanning for changes in some outcomes up to and then after some discrete event, like

when an intervention or new procedure was implemented.

Some resources about it are:

- [*Statistical Process Control*], a supplement to Polit & Beck (2017)





## **Part II**

# **Introduction to Psychometrics**



## Chapter 12

# NURS 925: Psychometrics Course

This “chapter” contains links to the presentations and materials covered in the psychometric course of the curriculum, NURS 925.

### 12.1 Foundations of Modern Research Measurement

- [Slides](#)
- This presentation covers:
  - Theories and domain sampling
  - Types of measurement
  - A brief history of measurement
  - General concepts of good measurement
  - Likert response formats

### 12.2 Validity and Reliability

- [Slides](#)
- Note: This presentation is a PDF document
- This presentation covers:
  - Historical conceptions of validity
  - Modern conception of validity
  - Reliability as measurement of true score
  - Types and measurements of reliability

### 12.3 Introduction to Factor Analysis & Exploratory Factor Analysis

- [Slides](#)

- This presentation covers:
  - Introduction to Factor Analysis
    - \* An Aside About Principal Component Analysis
  - Exploratory Factor Analysis
    - \* Concept of EFA
    - \* General Steps to Conducting EFAs

## 12.4 Confirmatory Factor Analysis

- [Slides](#)
- This presentation covers:
  - Review of Factor Analysis
    - \* EFA vs. CFA
  - Overall & Particulars of Conducting CFAs
  - Testing CFA Model Fit
  - A Few Words About Sample Size



## Chapter 13

# Exploratory Factor Analysis

### 13.1 The Concept of Factor Analysis

Factor analysis is a member of a rather large family of analyses that—among other things—uses ostensible variables to measure and analyze non-ostensible constructs, domains, or what are generally called factors. In this sense, a “factor” is that non-ostensible thing that determines whatever value ostensible variables take on. Factor is intentionally a very general term here that can encompass true constructs/domains, but is also simply whatever underlying thing is driving what we actually see.

It is *not*, however, intended to imply a factor in a statistical model, nor a “factor” in math (a multiplicative product or common divisor). So, please forget whatever definition you already have for the word “factor” and understand it here as simply that non-ostensible thing that drives one or more ostensible variables. (Indeed, one of the two main types of factor analysis, “exploratory factor analysis,” revolves around trying to figure out just what the underlying factors really are. I won’t discuss here the other main type of factor analysis, confirmatory factor analysis.)

#### 13.1.1 Role of Correlation/Covariance Matrix in Factor Analysis

All factor analyses per se begin with the assumption that the more that ostensible variables correlate with each other, the more likely they are to measure the same, underlying (non-ostensible) factor. This assumption may not be true, but factor analyses<sup>1</sup> assume it is. And what factor analysis does—in essence—is group variables together based on how well they correlate<sup>2</sup>; in fact, most statistical software can conduct a factor analysis on a correlation matrix of data alone: we don’t need to have access to the raw data (except to know the sample size). Given this, it may help to look at factor analysis first through the lens of a sample correlation matrix:

	Item 1	Item 2	Item 3	Item 4
Item 1	1	.80	.10	.20
Item 2	.80	1	.05	.15

<sup>1</sup>Other analyses related to factor analysis don’t necessarily make this assumption, and this assumption can be relaxed with, e.g., confirmatory factor analysis. Nonetheless, it may help to understand the basics by going with this otherwise common assumption.

<sup>2</sup>Remind you of the classical measurement theory’s concept of reliability?

	Item 1	Item 2	Item 3	Item 4
Item 3	.10	.05	1	.90
Item 4	.20	.15	.90	1

In this correlation matrix, Items 1 and 2 are strongly correlated with each other, but neither correlates well with Items 3 or 4. Alternatively, Items 3 and 4 correlated well with each other (but not with Items 1 & 2). In this example, we wouldn't be surprised if Items 1 and 2 measured the same thing and if this "thing" was largely unrelated to whatever Items 3 and 4 measured. In this example, then, we would expect that Items 1 and 2 are both ostensible manifestations of the same non-ostensible factor, and that Items 3 and 4 are the manifestations of an other non-ostensible factor.

With such a simple and clear example, we wouldn't need to conduct a factor analysis: Our eyes can be trusted well enough here. Often, however, the picture is not as clear or simple, and we may yearn for some objective process firmly grounded in common, sensible assumptions and well-tested by research. For these occasions, we may turn to factor analysis.

### 13.1.2 Factor Analysis Is Similar to the Linear Regression Analyses You Already (Should) Know

As I noted above, factor analysis itself relies on the correlation matrix (or, similarly, the variance/covariance matrix<sup>3</sup>). There are different ways to analyze this matrix (or derive it and related statistics from the raw data), but in general, factor analysis conducts analyses conceptually similar to multivariate multiple regressions<sup>4</sup> to determine how the items "load" onto the factors very similarly to how we estimate the *beta*- or *b*-weights for parameters in a linear regression. And like, e.g., an ANOVA, we not only get stats for how well our terms explain the data, we get stats (like Mean Square Error) for the extent to which our terms *don't* fit the data. (More on this much later.)

## 13.2 Steps to Conducting an Exploratory Factor Analysis

(Please note that I broke out the steps here a bit differently than I did in the presentation. The steps are the same in both, I simply divided the same procedure into a different number of steps based on how well each worked for the two media. Costello and Osborne (2005) offer more good guidance.)

### 13.2.1 1. Estimate the Number of Factors

In multivariate analyses like a MANOVA, we know ahead of time how many DVs there are. Similarly, to get measures of how well the ostensible variables (e.g., items<sup>5</sup>) load onto the factor(s),

<sup>3</sup>A variance/covariance matrix—also simply called a covariance matrix for short—is simply a correlation matrix before the values in it are standardized. (Remember, correlations are measures of shared variance—that are standardized to values between 0 and 1 (and designated as positive or negative depending on the direction of the values).)

<sup>4</sup>"Multivariate" here means that there are more than one criteria ("DV's") and "multiple" means there's more than one predictors ("IV's"); the criteria here are the non-ostensible factors we're estimating, and the predictors are the ostensible variables (e.g., items).

<sup>5</sup>So far, I've talked about the items that are analyzed for their inter-correlations. However, we could include any ostensible variable—not just the items of an instrument. We could include demographic variables, scores on other

we need first to get a sense of how many factors there may be. Therefore, the first step to an exploratory factor analysis (EFA) is to estimate the number of factors to use for further analysis.

The number of factors we choose could range from 1 to the number of items we have.

If we were to choose that the number of factors equaled the number of items (e.g., saying that there are 10 factors that underlie a 10-item instrument), we would essentially be saying that every item measures something different. This may be so, and would mean that there really isn't any reason to conduct a factor analysis since it wouldn't help us understand the data any better than item-level scores.

Often, though, when we conduct item-level analyses means we can't easily (and unbiasedly) tease apart interesting from uninteresting variance—or we may find all variance “interesting” and never see the forest for the trees. There are certainly times to scrutinize each item, but our ultimate goal is hopefully bigger—more profound—things than that.

Therefore, we strive to choose a smaller number of factors than items. But how much smaller? The general strategy is to find a number of factors that explains “enough” of the data. And so, a lot of thought has gone into what we mean by “enough.” One common criterion is to choose any factor that accounts for more of the data than an individual item does.

Another common criterion is to choose those factors that seem to “stand apart” from all of the other factors<sup>6</sup>.

### 13.2.1.1 1.1. Drop-offs in Scree Plots

One of the earliest and biggest contributors to factor analysis was Raymond Cattell. An astoundingly productive and brilliant man, Cattell was also not above controversy for what were either his political beliefs or misunderstandings of them<sup>7</sup>. He may or may not have believed—as Galton certainly did—in eugenics and other right-wing supremacist positions throughout much of his life. Whoever he was as a person, he was a gifted researcher, able indeed to see the “forest” and even to devise ways for us to do so better ourselves.

Perhaps his most lasting strategy was to use what he called “scree” plots to help decode how many factors to choose. “Scree” is the collection of rubble that gathers at the bottom of a cliff or plateau. The idea is that we're interested in the solid, prominent cliff and not the rubble. We can choose whichever factors “stand out” above the scree of other factors. In the following plot, we may thus decide that this 12-item instrument contains two factors that account for most of the variance since the first two dots make what nearly looks like a ledge with all of the other ten dots far below it—like rubble at the base of a cliff<sup>7</sup>:

### 13.2.1.2 1.1. Eigenvalues Greater than 1

#### Meanings of “Eigenvalue”

measures, or even predictors (IVs) and criteria (DVs). When we include predictors and criteria in our “factor analyses,” we are simply starting to turn them into structural equation models.

<sup>6</sup>There are other criteria one can use to determine how many factors to keep, including computing a  $\chi^2$  based on values from the matrix of residuals (i.e., what's left over in the variance/covariance matrix after running the factor analysis), the number of items, and the sample size. But, whatever criteria we use, it's never seen as meaningful to retain any factor whose eigenvalue is less than 1.

<sup>7</sup>I simply made that figure in an spreadsheet program, so I just made up the item loadings. If you're curious, their actual values are: 3.0, 2.7, 1.3, 0.8, 0.7, 0.65, 0.6, 0.55, 0.5, 0.45, 0.4, and 0.35 (which do indeed add up to 12).

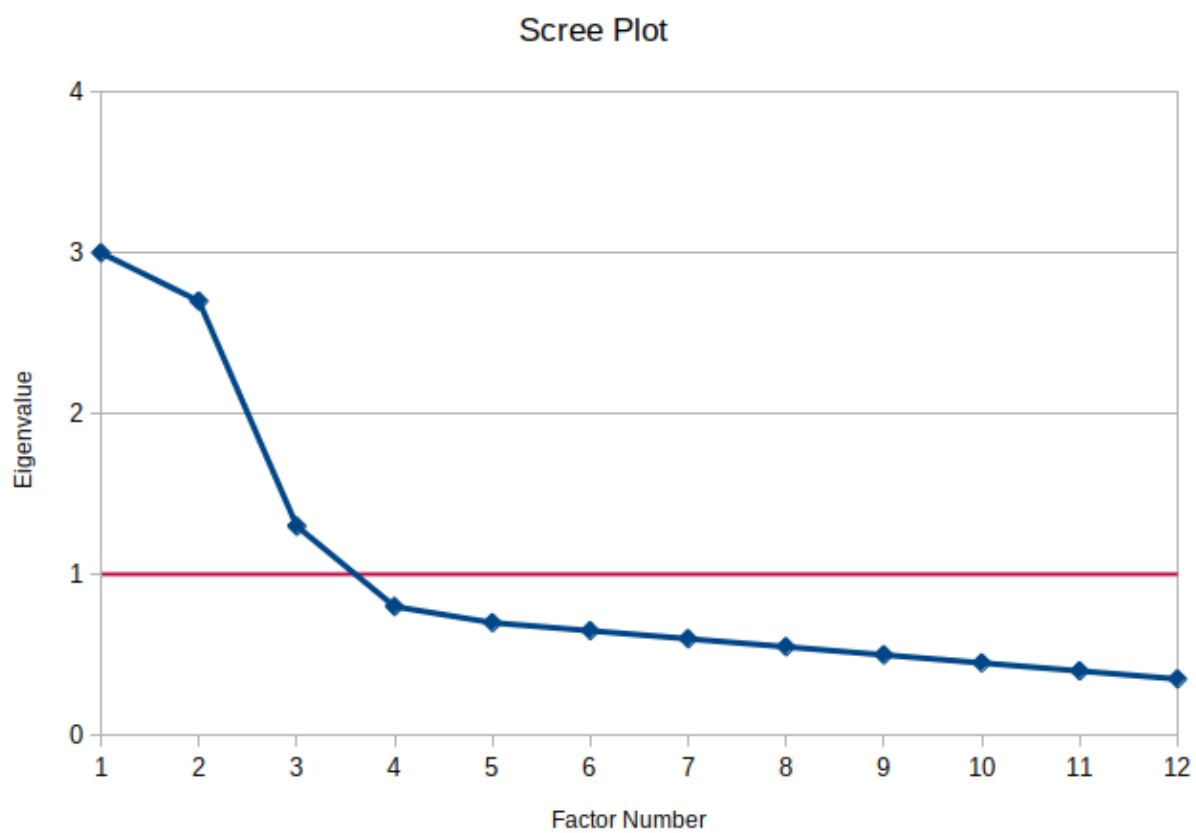


Figure 13.1: Example Scree Plot of a 12-Item Instrument

**Meaning 1: Proportion of Total Variance**

The  $y$ -axis in the scree plot measures each factor's **eigenvalue**. Eigenvalues, commonly used in mathematics and engineering, represent the amount of total variance explained by a given factor in the context of factor analysis.

Here's how it works:

- The **sum of all eigenvalues** equals the total number of variables (or items) analyzed. For example, in a 12-item analysis, the eigenvalues in the scree plot will sum to 12.
- The eigenvalue for the first factor is 3. This means:

$$\frac{3}{12} = 0.25$$

Thus, the first factor accounts for **25% of the total variance**.

If we compare this to correlations, the square root of this proportion ( $\sqrt{0.25} = 0.5$ ) suggests an average correlation ( $r$ ) of 0.5 between the items and the factor. This aligns with the principle that **variance is the square of correlation**.

For the first two factors (eigenvalues of 3 and 2.7), their combined contribution is:

$$\frac{3 + 2.7}{12} = \frac{5.7}{12} = 0.475$$

or nearly half of the total variance.

**Meaning 2: Sum of Items' Squared Loadings on a Factor**

An eigenvalue can also be understood as the **sum of the squared loadings of all items on a factor**. To illustrate:

Suppose a 4-item instrument has the following factor loadings:

$$\text{Factor}_1 = .7x_1 + .5x_2 + .1x_3 + .2x_4$$

Here: -  $x_1$  through  $x_4$  are scores on the items in the instrument. - Factor loadings (0.7, 0.5, etc.): The correlations between the items and the factor.

The eigenvalue for the factor is calculated as the sum of the squared loadings:

$$.7^2 + .5^2 + .1^2 + .2^2 = .49 + .25 + .01 + .04 = .79$$

Thus, the eigenvalue is 0.79. This represents the total variance accounted for by the factor across these four items.

For the 12-item scree plot example, the eigenvalue for the first factor is 3. This implies that the sum of the squared loadings of the 12 items onto the first factor equals 3.

**Why squared loadings?**

- Factor loadings are essentially correlations, and squaring them converts correlations into proportions of explained variance. Since an eigenvalue represents the **proportion of total variance accounted for by a factor**, it must be based on the squared loadings.

### Meaning 3: Relative Contributions of Individual Items

As noted earlier, the sum of all eigenvalues equals the total number of items. In the 12-item example, the eigenvalues sum to 12. This is because, by definition, the eigenvalue for any single item (if treated as a factor) equals 1.

- **Implication:** A factor with an eigenvalue greater than 1 accounts for more variance than any single item.
- Conversely, a factor with an eigenvalue less than 1 contributes less variance than a single item.

This is why a common criterion for factor retention is **Kaiser's Criterion**<sup>8</sup>, which retains only factors with eigenvalues greater than 1. Looking at the scree plot:

- Factors with eigenvalues less than 1 (e.g., Factors 4 through 12) do not contribute enough variance to justify their inclusion.
- Retaining such factors would mean adding complexity without meaningful additional information.

**Takeaway:** If a factor has an eigenvalue less than 1, it is generally better to rely on the individual items rather than attempting to interpret that factor.

#### 13.2.1.3 1.3. Theory (or Practicality)

I purposely made Factor 3 in that scree plot above just a little greater than 1 (it's 1.3). Since it's greater than 1, we may want to keep it. Since it's much lower than the next-larger factor (Factor 2), we may want to exclude it. We could use either criterion to justify our decision, so how do we decide?

Factor analysis—especially exploratory factor analysis—is not entirely an objective task. We should thus not simply swallow whole the results of factor analysis; we should chew on it and see if it tastes like something real. In this case, we could look at the results we get from choosing a 2-factor solution with those of a 3-factor solution and decide ourselves which appears to make more theoretical or practical sense.

#### 13.2.2 2. Evaluate the Results

The second conceptual step in exploratory factor analysis is to evaluate the results we obtain given the number of factors we chose. And yes, it's often worth playing around here, choosing different numbers of factors, and maybe even selecting subsets of items (as I did with when I removed the items from the APT that compared animals to humans).

In exploratory factor analysis, the main way we evaluate the results is by reviewing how well various items load onto the various factors. We review them to see if the factors make sense, if they provide any good insights, etc.

---

<sup>8</sup>It's called "Kaiser's criterion" after the research who first espoused it, Henry Kaiser, in Kaiser, H. F. (1974). [An index of factorial simplicity](#). *Psychometrika*, 39, 31 – 36; a clarifying review of that seminal article is [here](#).

Often when we review how well the items load onto the factors, we will find that items load pretty well onto more than one factor. To turn again to the 4-item example I gave in the presentation, the items had the following loadings onto the two factors:

$$\text{Factor1} = .7x_1 + .5x_2 + .1x_3 + .2x_4$$

$$\text{Factor2} = .1x_1 + .1x_2 + .6x_3 + .8x_4$$

Sure, the loadings of .1 are really small. But what about that item 4 that loaded a bit onto both factors?

One criterion is to simply choose the highest loading (putting item 4 onto Factor 2). **Another is to place items onto any and all factors on which its loading is >.3.**

Note that an item may not load well onto *any* factor. (Again, commonly this means that its loadings are all less than .3 on all factors.) Such items deserve especial attention—and may well be removed from all factors (and possibly even instrument scores and other analyses).

### 13.2.3 3. Review Different “Rotations” of the Factors

As I wrote just above, the four items load a bit onto both factors. Since all four item scores would be part of both factors’ scores, these two factors themselves would be slightly correlated. When factors are allowed to be correlated like this, we say they are “oblique” to each other; they are called oblique because if I plotted them as two axes, those axes would be at right angles to each; the term for two lines (axes) that are neither perpendicular nor parallel is “oblique.”

Why go through all the trouble of calling them oblique when we could just say they’re correlated? Because we can do further mathemagic on the factors and force them to be uncorrelated—and this involves “rotating” the axes so that they are indeed perpendicular to each other. Of course, even that is not eldritch enough, so instead of saying the factors are now perpendicular, we say they’re “orthogonal,” which means the same thing.

To make matter even worse, there are different ways to rotate factors into orthogonality—and even different ways to rotate them to be variously oblique. The relative merits of the growing number of rotation techniques is a [vibrant area of research](#), with [no clear winner found](#).

#### 13.2.3.1 Choosing Which Rotations to Use

In general, it’s [good advise](#) to try out at least one orthogonal and one oblique rotation and see how this helps us interpret the data.

#### Orthogonal Rotations

Orthogonal rotations force all factors to be unrelated to each other. These rotations therefore use various procedures to maximize the differences between factors.

- **Varimax** is often the default choice when there is no strong preference for another method. Its primary goal is to simplify the interpretation of **factors** by redistributing the variance of loadings to make high loadings higher and low loadings lower. Specifically, it maximizes the variance of the squared loadings for each factor across all items. In practice, this results in a factor structure where each factor has a few strongly loading items and many items with near-zero loadings. This simplification helps make the factors more interpretable. Varimax does not inherently make factor eigenvalues more similar or the proportion of explained

variance more equal among factors, though it can result in a more balanced distribution in some cases.

- **Quartimax** is another common orthogonal rotation method. While varimax aims to simplify the interpretation of *factors*, quartimax focuses on simplifying the interpretation of **items**. It does this by maximizing the variance of squared loadings for each item across all factors. In practice, quartimax tends to produce a factor structure where items load strongly on a single factor while having minimal loadings on others. This can make it easier to identify which items belong to which factor. However, quartimax often results in a factor structure where the first factor dominates, accounting for the largest proportion of the total variance. This makes the factors less balanced in terms of variance explained. For example, applying quartimax to the first two factors in a scree plot would likely accentuate the difference in eigenvalues between those factors, making the first factor seem more dominant.

There are plenty of other orthogonal rotations, but both of these have stood the test of time (and even some formal scientific tests) and should work well enough for most data.

### Oblique Rotations

There are even more oblique rotations. They also tend to require a bit more hands-on involvement from the researcher. Luckily, though, for the two I describe next, this just means deciding the maximum possible value for how strongly the factors can correlate with each other. In other words, I could determine ahead of time that I want my factors to have correlation coefficients with each other that are no larger than .4.

- **Promax** is a common oblique rotation that in fact begins with an orthogonal rotation before “relaxing” this orthogonality requirement enough to let factors correlate with each other (either to a least-squares sense or to the maximum level you set ahead of time.) It tends to give rather easily-interpreted loadings onto the factors.
- **Direct Oblimin** doesn’t begin with an orthogonal rotation like promax but instead attempts to find the best axis for each factor and then modifies this slightly to try to reduce how much items load onto each factor.

Promax will tend to produce factors that are more orthogonal than direct oblmin.

### 13.2.4 3. Interpretation

This is arguably not a single, discrete step. Instead, one should be considering what is happening at each of the other steps, reflecting upon how it relates to apposite theories, and using one’s own judgment to guide analyses.

Nonetheless, researchers do tend to step back at the end of one round of factor analysis to consider how things look. I, too, recommend making sure there is at least this one deliberate pause to reflect on what the results imply about data and their larger meanings.

Similarly, you may now want to stop to think about all that I’ve written here—and let me know what questions you have.





## **Part III**

# **Guides to Using Software**



## Chapter 14

# Using Templates and a Reference Manager

*Work smarter, not harder* –Scrooge McDuck et al.

### 14.1 Overview

This guide covers how to install and use both a style template and reference manager with Microsoft (MS) Word (after a brief description of their use with LibreOffice Writer).

### 14.2 Style Template

The idea behind style templates is that you separate out the plain-text content of a file from the way it is formatted, including the typeface/font, colors, how phrases are emphasized (e.g., with *italics*), etc. Separating out these two things greatly helps documents move from one platform or program to another. It also helps if, e.g., you submit a manuscript to one journal that uses APA format, but it gets rejected and you want to then submit it to another journal that uses, say, MLA. With templates, you simply choose a new template; without templates, you spend your afternoon plodding through your manuscript re-formatting headers, redoing every citation, etc. Using style templates thus help write because it lets you not worry about how to even put it into, say, APA format in the first place.

#### 14.2.1 Loading Templates in MS Word

You can [create your own](#) template<sup>1</sup>, but it's nearly always easier to use an existing one. There are several APA templates available online, including an “official” one from [Microsoft](#)<sup>2</sup>, ones from [MS proponents](#), from [coders](#), and from [other schools](#). The one we'll use is available [here](#).

---

<sup>1</sup>And the one we'll use is one that I created.

<sup>2</sup>Although MS offers none for the current edition of the APA Manual, and they have not put much effort into making such templates complete or ensuring their accuracy. After all, it's hard to support both another billionaire CEO *and* quality products!

### 14.2.1.1 Loading a Template for the First Time

1. Download the [APA\\_7th\\_Ed\\_Template.dotx](#) Word template to your computer.
2. Open this file in MS Word.
3. Choose to Save As the file
  1. Near the bottom of the dialogue box that opens, under Save as type, select to save it as a Word Template.
  2. Word should automatically save it in the Documents > Custom Office Templates folder. This will allow you to more easily access this template in the future (as we'll discuss just below)

Note that you can now use the template to write your paper. Now, though, when you save it, save it as a normal .docx file—**not** a .dotx template. (You *can* save it as a .dotx file and use that, but Word treats them differently and this will create a bunch of un-needed templates.)

### 14.2.1.2 Loading an Existing Template

Now that you've loaded the template into Word, you can more easily reuse it. Let's start from the top to do this, so first close that template file and Word. Now, re-open Word.

#### From the Opening Dialogue

1. Word, of course, opens to a dialogue that lets you choose recent files from the left-hand menu or to open a new file, using any of the mostly-useless templates that it gives in the main window on the right.
2. You most likely only ever chose to open a Blank document, but now you should see the APA template you saved listed among all of the other templates.

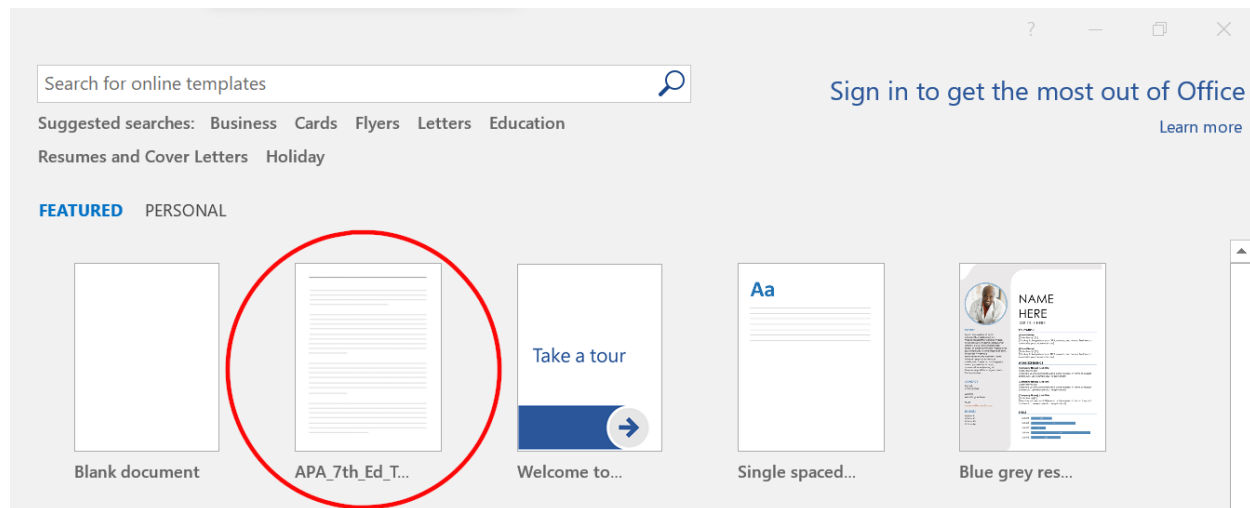


Figure 14.1: Opening the Template You Added to Word

3. If you don't immediately see it, click on the word Personal next to the currently-highlighted Featured word.

4. You should see your APA template there—perhaps even as the only Personal template.

There is more to adding, opening, and modifying templates. [TechRepublic](#) has a good coverage of that, so we'll jump now to using them.

### 14.2.2 Loading Templates in LO Writer

[LibreOffice Writer](#)—a free alternative to MS Word—works well with style templates.

Although there is an [extension](#) you can use to apply a style template, this is easily done by:

1. Open a template like you would any other file in LO Writer. A template for the 7th edition of the APA *Publication Manual* is [here](#).
2. Click on Insert > Document<sup>3</sup>.
3. Select the existing document you want to apply the style to.
4. Save As a new version of the document that's either an ODF Text Document (.odt) or Word 2007–365 (.docs) file. I.e., that's not a template (.ott) file.

You can most easily change the style template being used with the [Template Changer extension](#):

1. Download & load the extension
  1. [Download the Template Changer extension](#)
  2. In LO Writer, click on Tools > Extension Manager... (Cntrl + Alt + E)
  3. In the dialogue that opens, choose to Add, and then navigate to where you downloaded the extension
  4. Select and Open the downloaded extension (or, of course, just double-click on it)
  5. Accept the license agreement (if you indeed do)
  6. Restart LO Writer
2. Click File > Templates > Change template (current document)...

### 14.2.3 Using a Template

#### 14.2.3.1 MS Word

The template we're using already has the outline for a typical article manuscript (and a few other things we'll discuss later in this guide). You really simply need to start typing, and then Save as a normal .docx file to use it. There are a few things to point out and practice, though.

Again, you *can* use a template to do things like italicize words, but the main use will be for the section headings and sub-headings. To use the style template:

1. Under the Home ribbon, notice the Styles section
2. In the lower, right corner of the Styles section, click on the tiny box-with-arrow button to expand that section...

into a menu that defaults onto the right of the Word window:

---

<sup>3</sup>In the Ubuntu version of Writer, the menu selections are Insert > Text from File....

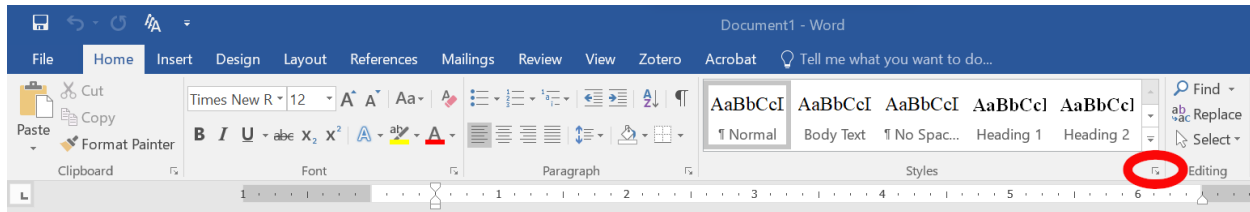


Figure 14.2: Opening the Styles Menu

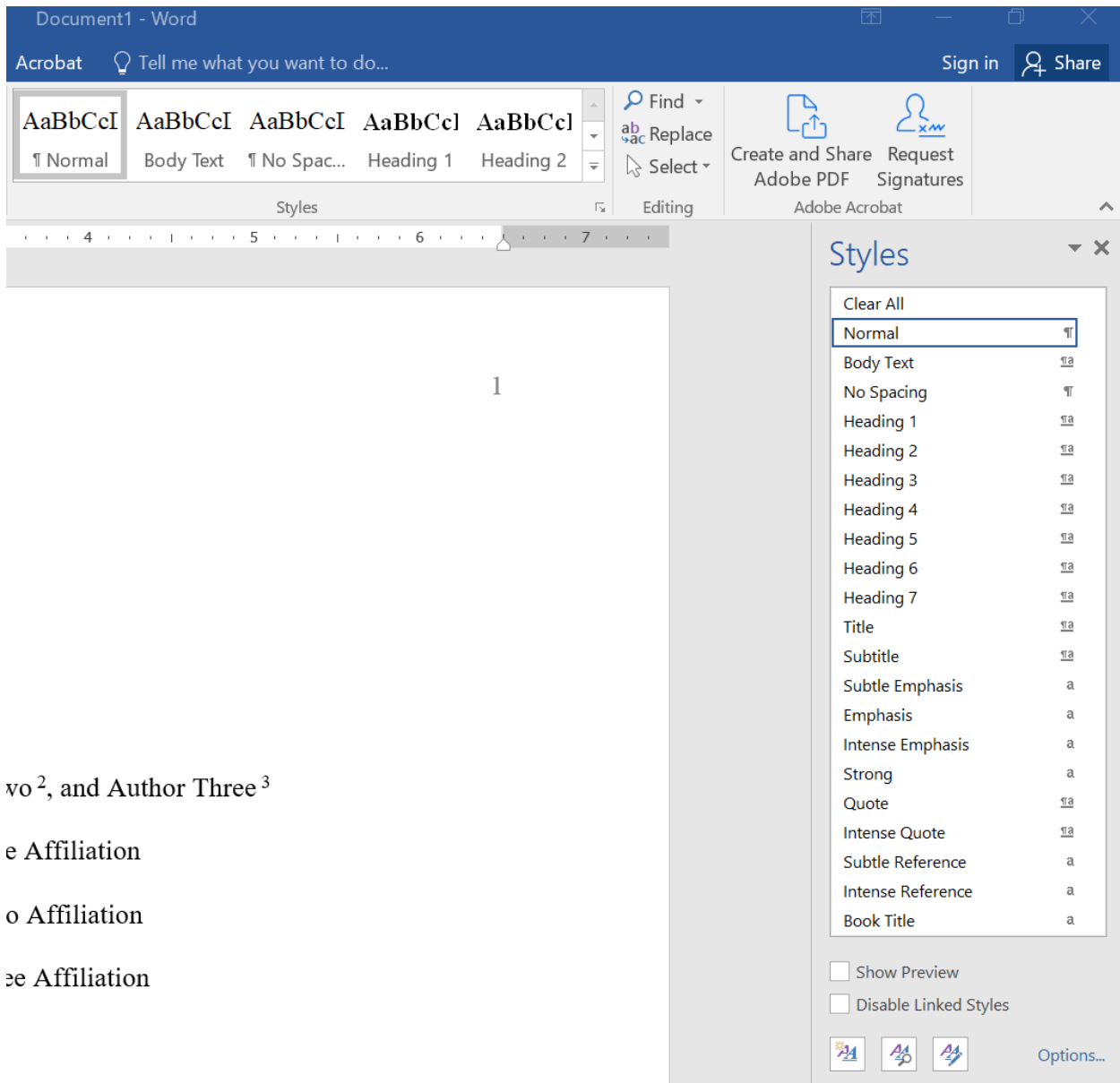


Figure 14.3: The Styles Menu

You can now more easily use the styles.

I labelled the headings levels in the template<sup>4</sup>, but you can also tell what formatting is being applied to a given section by left-clicking on a piece of text; the formatting that's being applied will be highlighted in that right-hand menu.

Clicking, e.g., on the document's **Title** shows that it's being formatted as a Level 1 heading. If you click on, say, Heading 2 in that menu, you will see that the section that's currently chosen will change its formatting to be like the other Level 2 headings (like **Participants** is formatted under the **Methods** section).

Normal text also has its own style. Clicking on the normal text under **Descriptive Statistics** in the **Results** section shows that it's formatted as Body Text. We could have used the Normal formatting just above Body Text in the menu; it really makes no difference—as long as both of those style elements are formatted the same (as they are here).

You can easily switch between style elements by clicking on different elements in the styles menu for a given piece of text. Note that headings and title elements in the menu will be applied to an entire paragraph<sup>5</sup> while most other elements can be applied to single words (or other parts of a paragraph); there is a small pilcrow (¶) next to paragraphs styles in that menu and a small letter a next to non-paragraph styles.

(The Figure 1 title in the template is styled as Strong to show you how you *could* use styles to even make bold and italics text, even though you'll likely just use Control/Command + B and Control/Command + I instead.)

Note that you can change the styling of the style elements (e.g., change how Level 1 headers look). To do this, simply right-click on an element in that styles menu and select Modify.... You can change the font elements, or—from the Format button at the bottom of the dialogue that opens—change the styling of the whole paragraph (or other things like the page border). When you open this Modify Style dialogue, note that you can apply this modification to just this file or to the template itself for use in other files:

Among the changes you may want to make to this template are taking out the sub-headings under **Participants**. I put them there simply to show all of the heading levels that APA allows, what they look like, and where they're located in the styles menu.

#### Note

When using a template to write a manuscript, remember to Save as a normal .docx file once you've started using. This will keep the styling but not screw up your template file. If you want to make changes to the template, then Save as a .dotx file instead.

## Applying a Template to an Existing Document in Word

It is relatively easy to apply a new template to an existing document—even if Word doesn't make the steps to do so intuitive.

There are indeed times when you may want to apply a new template to an existing document. Many journals, for example, require you to format your submitted manuscript into their own

<sup>4</sup>Once you're more familiar with doing this, you may want to delete those extra cues—and the extra text—and re-save this as a cleaner .dotx template file.

<sup>5</sup>This actually presents a problem for APA style. APA dictates that level 4 and 5 headings show not be on their own line, but simply at the beginning of a paragraph, with the rest of the paragraph formatted like normal text. I don't know how to get around that without some kludgy elements that, say, have negative margins on the bottom.

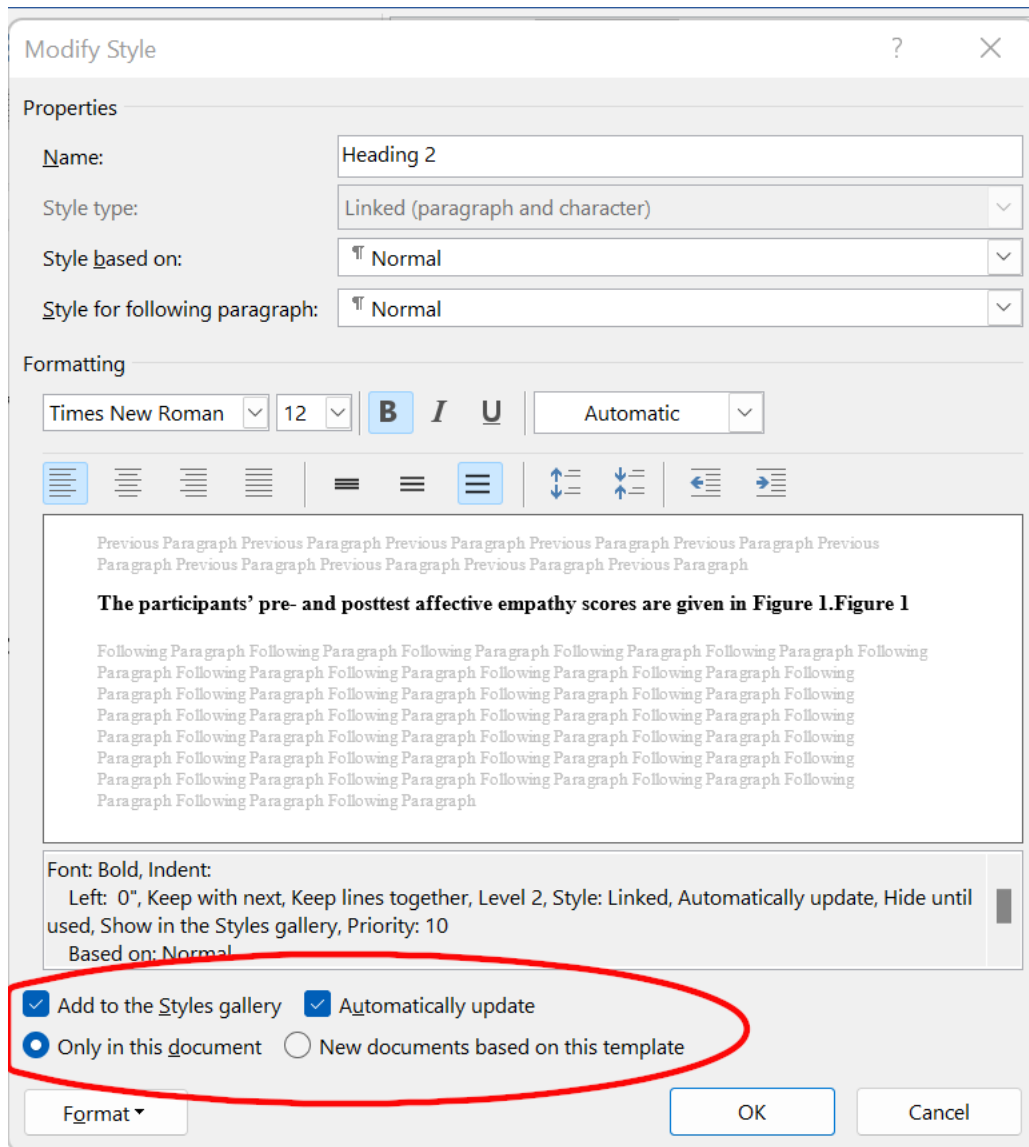


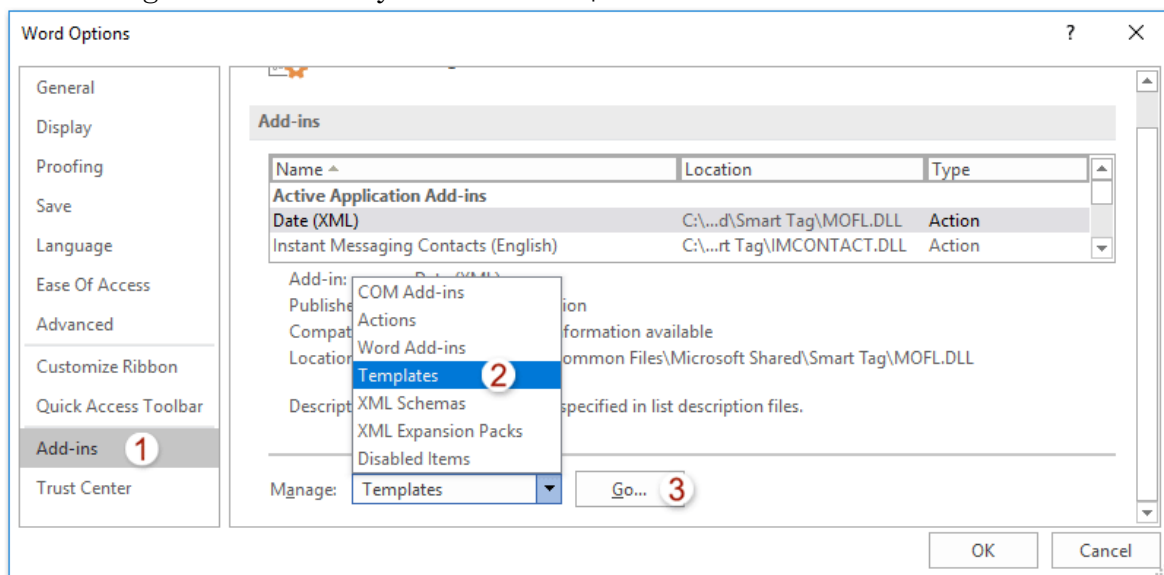
Figure 14.4: Applying style element modifications to either just this file or to the template itself

style. So, if you submit a manuscript for publication to one journal, have it rejected, and then apply to an other journal, you will need to reformat the entire document for the second journal. This is a *lot* easier to do by simply applying a new style template to it.

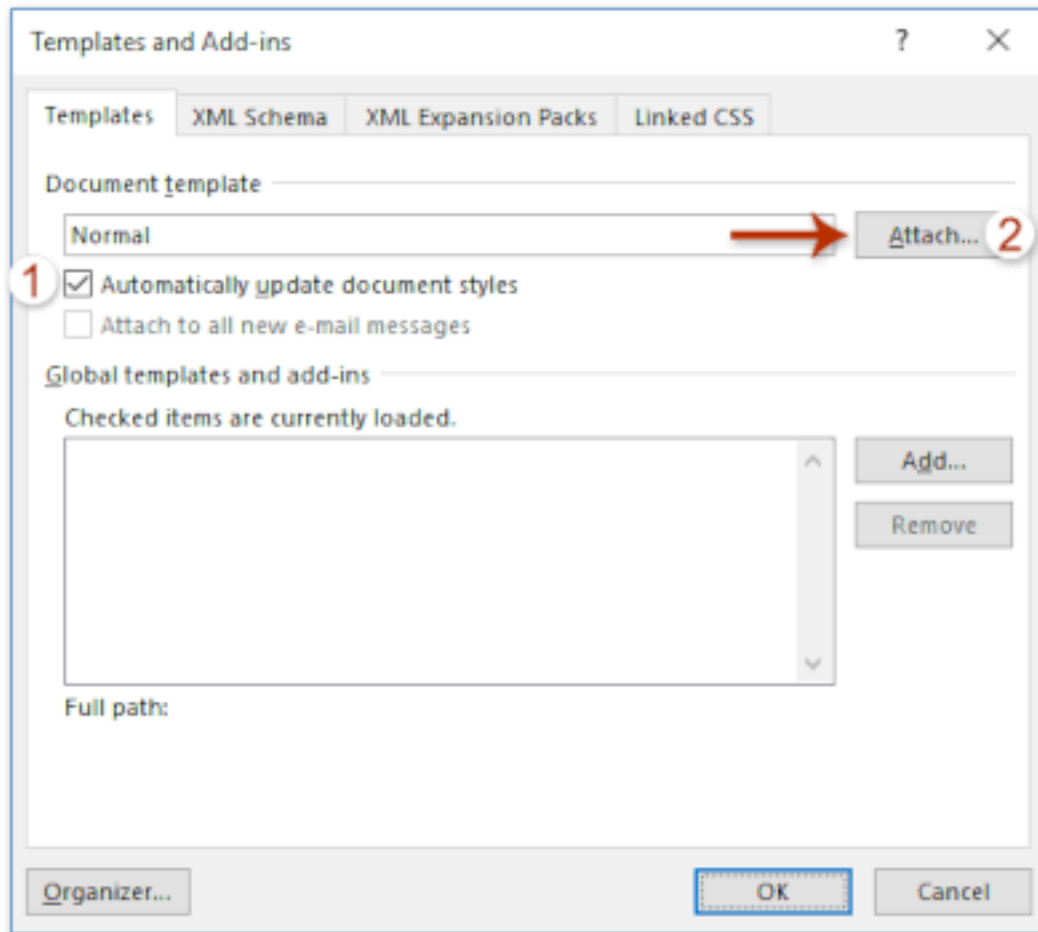
Please note that this is more useful if the existing document you're applying the new template already used styles to format the headings, etc. I covered how to do this above in Section 14.2.3.1.

The steps to doing so for MS Word are [here](#). To summarize those steps (while only stealing a few of their nice graphics):

1. Open your existing document in MS Word
2. Click on File > Options.
3. In the dialogue box that opens, click on Add-ins near the bottom of the menu on the left. This is step 1 in the image below.
4. In the main part of that dialogue box, there is a Manage section near the bottom. From that, choose Templates.
5. Click Go right next to where you selected Templates.



6. In the new dialogue box that opens, *make sure to select* Automatically update document styles.



7. Click Attach and then OK in that same dialogue.
8. Navigate to the new style template (i.e., the .dotx file) you wish to use. Select it, and click Open at the bottom.
9. Click OK when you return to the previous dialogue (shown in Templates and Add-ins dialogue).
10. You can now (re)save your document with the new style applied.

#### 14.2.3.2 LO Writer

1. Open a style template file in Writer. Templates' extension is .ott; an example of one is [this APA 7th Edition template](#).
2. Save the file as an ODF Text Document (.odt) file; this is the default extension for all open document text files, like those created by LO Writer.
3. There are three ways to change styles of parts of the document:
  1. Click on the Styles drop-down menu
  2. Click on the Styles drop-down menu that may appear in the the toolbar.
    - This option may not be added by default, depending on the version of Writer, but can be easily added via Tools > Customize... > Toolbars where it can be added to one of the toolbars.
  3. Tapping F11 to open aside menu with all of the various types of format-able fields.
  4. Using the keyboard directly where, e.g.:

- Cntl + 1 (Command on a Mac) formats the paragraph as a level 1 header
- Cntl/Command + 2 formats as a level 2 header, etc.
- Cntl/Command + 0 formats as text body

I find using Cntl/Command + works best since I can simply type my manuscript and just hit, e.g., Cntl/Command + 1 and keep typing, say, “Methods,” hit Enter to start a new paragraph, hit Cntl/Command + 2 and type “Participants,” hit enter again, and then start writing that section. That’s it, with just a few key combinations to remember you can not worry about formatting.

And you may have noticed in what I wrote, most styles are applied to paragraphs, not—say—individual words or phrases inside a paragraph. You *can* use styles to do this (and sometimes I do), but just Cntl/Command + I to italics, Cntl/Command + B for bold, etc. work just as well; using those common key combinations will apply whatever style has been set for italics, bold-face, etc.

## 14.3 Reference Manager

Styling documents is one thing that computers *ought* to do well and relieve us from having to do ourselves. Managing our citations is certainly another. A reference manager allows you to collect and organize your citations. More importantly here, it lets you add them to your manuscript easily and to create your References section automatically. The only issue is if you collaborate with others on the same manuscript, you've got to use the same citation file format (.bib vs. .ris etc.)

I will cover two reference managers, RefWorks (Section 14.3.1) and Zotero (Section 14.3.2). RefWorks is useful for MS Word on Windows & Apple OSs; Zotero is good for GNU/Linux OSs. Please note that [other popular reference managers](#) include [Mendeley](#), [Endnote](#), and [RefWorks](#). I used to use [JabRef](#) and still kinda miss it.

### 14.3.1 RefWorks

RefWorks is a web-based reference management tool that helps researchers organize citations, generate bibliographies, and integrate seamlessly with word processors. This tutorial provides step-by-step instructions for:

- Downloading and setting up RefWorks (Section 14.3.1.1).
- Installing and using RefWorks Citation Manager (RCM) in Microsoft Word (Section 14.3.1.2).
- Using RefWorks with Google Docs (Section 14.3.1.3)
- Using RefWorks with LibreOffice Writer (Section 14.3.1.4).

#### 14.3.1.1 Downloading and Setting Up RefWorks

RefWorks is a web-based citation management tool, meaning it does not require installation on your computer. However, it does have browser plugins, word processor add-ons, and import/export tools that need setup.

#### Step 1: Creating a RefWorks Account

1. Go to the [RefWorks website](#).
2. Click on Create Account and go through the steps listed.
  - RefWorks asks for your institutional email; using this should let you use RefWorks for free. Hopefully you still can after you graduate.
3. Follow the registration prompts to set up your account.

#### Step 2: Installing Browser Extensions (Optional)

RefWorks offers browser extensions for importing citations from websites.

- **Chromium / Google Chrome:**
  - Install the Save to RefWorks extension from the [Chrome Web Store](#).
  - Click on the extension icon and log into your RefWorks account.

- **Firefox:**

- Install the Save to RefWorks add-on from the [Firefox Add-ons Store](#).
- Click on the extension and sign in.

### 14.3.1.2 Using RefWorks with Microsoft Word

RefWorks integrates with Microsoft Word via **RefWorks Citation Manager (RCM)**, available as an add-in.

#### Installing RefWorks Citation Manager (RCM) in Microsoft Word

##### Windows and macOS:

1. Open Microsoft Word.
2. Go to Insert > Get Add-ins.
3. Search for **RefWorks Citation Manager**.
4. Click Add to install.
5. Open the RefWorks Citation Manager from the References tab.
6. Log into your RefWorks account.

#### Using RefWorks in Microsoft Word

1. Open **RefWorks Citation Manager** in Word.
2. Select a citation from your RefWorks library.
3. Click Insert Citation to add it to your document.
4. To create a bibliography:
  - Click Bibliography Options > Create Bibliography.
  - Select a citation style (e.g., APA, MLA, Chicago).

### 14.3.1.3 Using RefWorks with Google Docs

#### Installing and Using RefWorks in Google Docs

1. Open **Google Docs**.
2. Click on Extensions > Add-ons > Get add-ons.
3. Search for **RefWorks** and install the **RefWorks Citation Manager** add-on.
4. Open the add-on by navigating to Extensions > RefWorks Citation Manager > Start.
5. Log into your RefWorks account.

#### Using RefWorks in Google Docs

1. Open **RefWorks Citation Manager** in Google Docs.
2. Select a citation from your RefWorks library.
3. Click Insert Citation to add it to your document.
4. To create a bibliography:
  - Click Bibliography Options > Create Bibliography.
  - Select a citation style (e.g., APA, MLA, Chicago).

### 14.3.1.4 Using RefWorks with LibreOffice Writer

RefWorks does not have a direct plugin for **LibreOffice Writer**, but you can still use it effectively.

#### Method 1: Manually Exporting Citations

1. In RefWorks, select the citations you want.
2. Click Export > Bibliographic Software.
3. Choose RIS Format and download the file.
4. Open LibreOffice Writer, insert references manually.

#### Method 2: Using the Write-N-Cite Tool

[Write-N-Cite](#) is available for **Windows and macOS**, but not for GNU/Linux. It allows inserting citations directly in **LibreOffice**.

1. Download Write-N-Cite from RefWorks ([link](#)).
2. Install the tool and log in.
3. In LibreOffice, open Write-N-Cite and insert citations.

#### Method 3: Using the Quick Cite Feature

1. In RefWorks, go to Cite > Quick Cite.
2. Select citations and copy-paste them into LibreOffice Writer.

### Troubleshooting and Additional Tips

#### Common Issues and Solutions

Issue	Solution
RefWorks add-in not appearing in Word	Restart Word and reinstall from the Add-ins store.
Save to RefWorks not working in browser	Clear cache and reinstall the extension.
Bibliography formatting issues	Ensure correct citation style is selected in RCM.
Linux users can't use RCM	Use <a href="#">Quick Cite</a> or export citations manually.

For more information, visit the [RefWorks Help Center](#).

### 14.3.2 Zotero

Zotero has a rather user-friendly set of guides in their [Documentation](#) section. I'll cover only the most common features now.

1. Go to the [Zotero Download](#) page where there should be shown options to install it into your given operating system.
  - Note that you can also install it into your browser from there. More about this later.
2. When you run the file you just downloaded, it will install Zotero on your computer.
  - It should also automatically integrate itself into Word. If you need to, though, you can [manually install](#) that Word functionality.

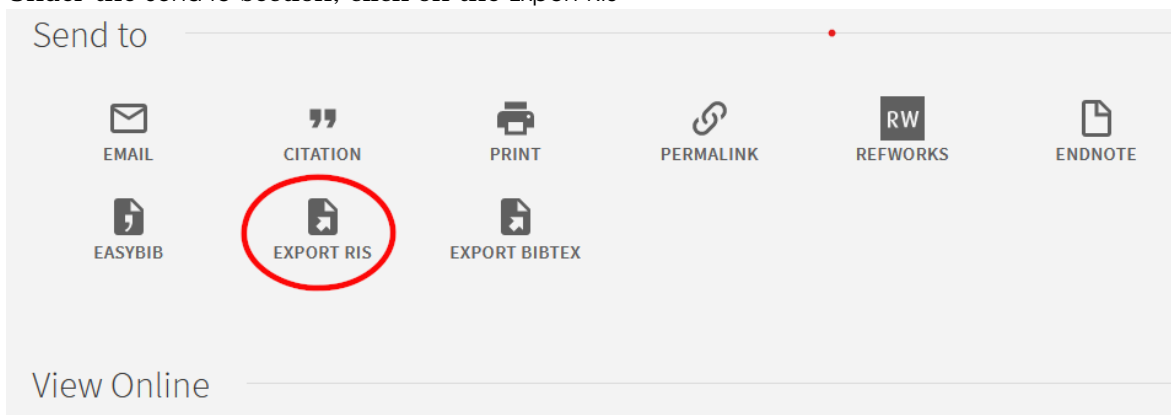
### 14.3.2.1 Adding Citations to Zotero

1. I've added a few sources into a file in BlackBoard as Sources to Add to Zotero, but you can also download that same file [here](#). Do please save that file to your computer.
2. In Zotero, click on File > Import... (or Control/Command + Shift + I)
3. In the dialogue box that opens, leave selected A file (BibTeX, RIS, Zotero RDF, etc.) and click Next
4. Navigate to the file you downloaded (it's named zotero\_sources.bib) and Open it
5. In the Options dialogue, I suggest unselecting Place imported collections and items into a new collection<sup>6</sup> before clicking Next then Finish

Ta-da! You have your first set of citations.

You can also get citations from, e.g., the Hunter library:

1. Go to the [main page](#) and enter something into the OneSearch field.
2. Click on a source you like to open the page about it.
3. Under the Send to section, click on the Export RIS



4. For Encoding, choose UTF-8 and then click Download and then Save it
5. You can now navigate to this file and import it just like you did that one I created for you.

Note that a good number of publishers will allow you to download articles and citations directly to your library. So, sometimes it's worth clicking to View Online in the Hunter library page for a source, go to the article's site under the publisher, and download it directly to Zotero from there.

If you've added the [browser extension](#), you can also use that both to access citations (through your Zotero account if you've set on up) or to download sources through that. This works rather well (in FireFox at least; I rarely use Chrome/Chromium or Safari).

<sup>6</sup>You can create several libraries, e.g., one for each line of research you pursue. However, I find that gets complicated and doesn't really help. So, I put all of my citations into one library.

### 14.3.2.2 Syncing Library

If you want, you can to [Register](#) for an account at Zotero. Doing this allows you to upload your library to their server and to use that to sync your library of sources across machines. There are options to pay for more storage, but I doubt you'll need to do that; I accidentally have over 13,000 sources in my library<sup>7</sup> and still haven't run out of room.

Once (if) you've set up an account, you can click on the [Web Library](#) link to access your citations online. You can also synchronize them with your local instance:

1. In Zotero on your computer, click on Edit > Preferences
2. Click on the Sync tab
3. Click to Link your account
  1. I suggest selecting to Sync automatically and perhaps to select all other options to syn full text, attachments, notes, etc.
4. Zotero will ask you to log in; with larger libraries, it can take several minutes to sync the first time, but subsequent syncs are as quick as any such operations with, e.g., Dropbox.

### 14.3.2.3 Using Zotero to Cite Sources in Word

#### Setting up Preferences

1. Make sure Zotero is open
2. In Word, open/create the file you'd like to import citations into
3. There *should* be a Zotero tab near the right end of your list of ribbons; if not, can [manually install](#) it.
4. Go to the place in the text where you want to insert the citation; leave the cursor there
5. Click on the Zotero tab to access that ribbon
6. Click on Document Preferences in that ribbon
7. In the dialogue that opens, choose the citation style you want to use from the Citation Style list
  - Unless you work with those like me who use LibreOffice, leave it to Store Citations as Fields and make sure Automatically update citations is also selected.

#### To Add a Citation in Text

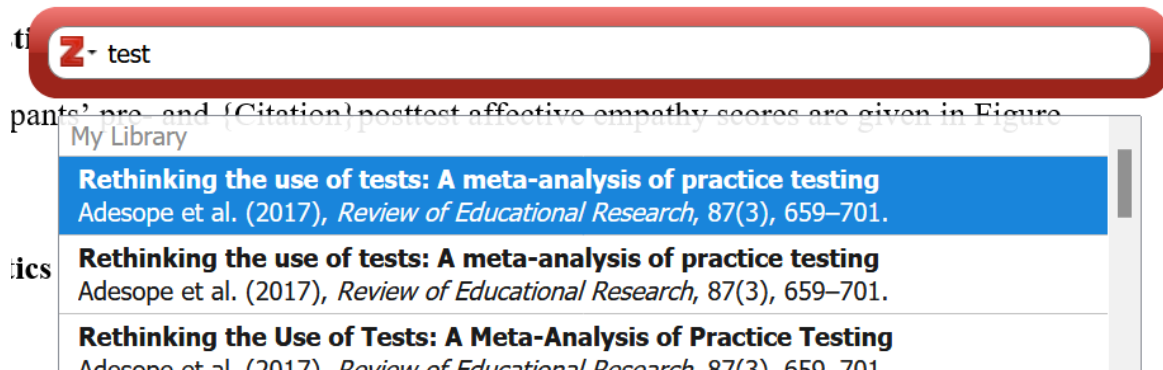
1. Make sure Zotero is open
2. In Word, open/create the file you'd like to import citations into
3. There *should* be a Zotero tab near the right end of your list of ribbons; if not, can [manually install](#) it.
4. Go to the place in the text where you want to insert the citation; leave the cursor there
5. Click on the Zotero tab to access that ribbon
6. Click on the Add/Edit Citation button

---

<sup>7</sup>Mistaken duplications

7. A very slim dialogue opens where you can search for the citation(s) to add:

### Results



- Note that you can search by author, title, journal, keyword, etc.
8. Select the citation you want from the drop-down list
    - If you want to add more citations, simply search for and choose them as well from that same dialogue
  9. You now have a few options for how to include that citation:
    1. If you leave it as it is and simply hit Enter, it will add a parenthetical citation, e.g., "(Cohen, 1988)"
    2. If you instead left-click on the citation, another dialogue will open:
      1. Selecting to Omit Author will add just the data in parentheses, e.g., "(1988)"; you would simply type in, e.g., "Cohen" before that to note the authors
      2. Selecting Page will allow you to add page numbers, e.g., "(Cohen, 1988, p. 200 – 201)"
      3. Prefix allows you to add text before the citation in the parentheses, e.g., "(see Cohen, 1988)"
      4. Suffix allows you to add text after it, e.g., "(Cohen, 1988, and others)"

### To Create a Reference Section

Simply click on the Refresh button in the Zotero ribbon. This will create and update your reference section with any citations you've added or removed. In fact, having selected Automatically update citations under Document Preferences should suffice.

#### 14.3.2.4 Creating Figure and Table Fields

One more thing to automate: the numbering of figures and tables. This is a bit kludgy in Word, but still worth it.

### Adding and Captioning a Figure or Table

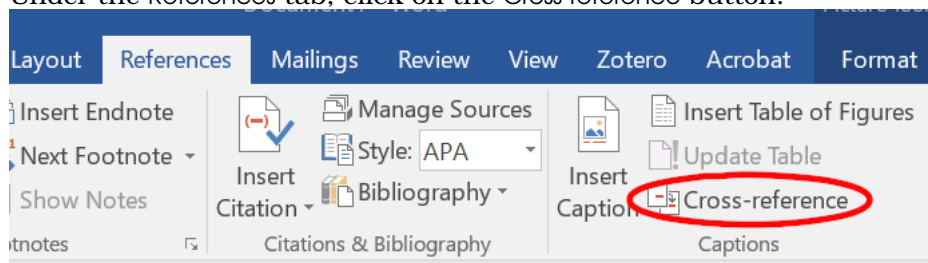
1. Add your figure or table
  1. Figures are usually inserted as simple image files (Insert > Pictures)

2. Tables are best done via:
  1. Insert > Table
  2. Use your mouse to drag your cursor to select the number of rows and columns you want (or choose Insert Table there to type in the number of columns & rows along with formatting of them)
2. Tables should automatically gain a caption, but figures won't. To add a caption for a figure:
  1. Right-click on the image
  2. Click on Insert Caption
  3. Under Options, choose Figure
  4. Under Caption, *maybe* type in a title for your figure. Note that this will put it on the same line as "Figure 1" which is not strictly APA; APA dictates adding the title on the line below "Figure 1," as I did in the template
  5. Under Position, choose Above selected item since that's what APA wants
  6. Select to Exclude label from caption
  7. Click OK

### Inserting Updatable References to Figures & Tables

1. Now, go to an other place in the manuscript where you want to reference that figure or table.

1. Under the References tab, click on the Cross-reference button:

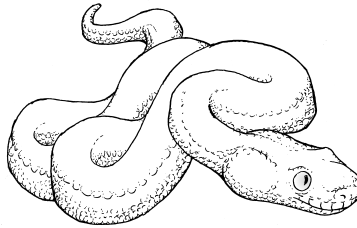


2. Under Reference type, choose Figure or Table as applies
  3. Set Insert reference to Entire caption (unless you didn't select to Exclude label from caption from the Caption dialogue or, of course, if you want to include it)
  4. Doing all of this should produce a list of the figures or tables in the For which caption field; simply choose the one you want.
2. Word will update the numbering of figures and tables . . . if you ask it to; it won't do it automatically. To update them:
    1. [As per these instructions](#), you can choose to update them by selecting text with similar formatting
    2. Or simply select all text (Control/Command + A)
    3. Right-click on the reference to a figure/table in the text and choose Update field from the drop-down menu that appears
    4. If a dialogue opes asking what to update ("Update Table or Figures?"), then choose Update entire table
    5. And click OK

That should do it. Again, it is a bit kludgy in Word. I included an example of such a cross reference in that template—yet another thing you may wish to delete before making that template your own.

## 14.4 Additional Resources

- Working with tables
  - [Putting tables in APA format](#)
  - [More manipulations of descriptive tables](#) including removing columns and rows and pivoting tables





## Chapter 15

# Introduction to Excel

There are times when I still prefer to use a spreadsheet program [instead of R](#) (or SPSS). Sure, it's easier for me in part because I've been using spreadsheets, well, pretty much since they existed<sup>1</sup>, but mostly because I like to look closely at the data, sometimes cell by cell, and to have that very precise and close-up control over it. I think you'll appreciate why it's such an enduring way to examine—and even analyze—your hard-earned data, especially after you master using the keyboard to navigate and sort data.

We will cover some of the most common functions and keyboard shortcuts. More keyboard shortcuts are given in [Microsoft's page](#) dedicated to that. Their site also has an [overview of formulas](#).

### 15.1 Overview

This activity is intended to review some of the main ways that you may use a spreadsheet program (viz., MS Excel) to manipulate, review, and even analyze data.

### 15.2 Filling in Series

1. Enter the first few values of a series in a row or column
2. Highlight those cells
3. Move cursor to bottom right corner of the last cell
  - The cursor should change shape (e.g., from a fat, white cursor to a skinny, black one)
4. Hold down the (left) mouse button
5. Drag the cursor right/down to fill in more values of that sequence

Excel (and [LO Calc](#), [Gnumeric](#), etc.) are pretty good at recognizing what the series should be. E.g., counting by even numbers, counting by tens.

---

<sup>1</sup>I miss Quattro Pro nearly as much as Word Perfect. Then again, I also miss making mixed tapes on my boom box and decorating the cassette cases, so take it for what it is.

## 15.3 Pasting Special & Transposing

- Data can be pasted with various levels of formatting retained or removed
- This is especially useful in two situations:

### 1. Converting formulas to raw numbers

1. (After creating some range of cells that contains formulas), select the range of cells you'd like to convert from formulas to raw numbers
2. Copy that range (Control + C)
3. Move to the range of cells you'd like to paste it to
  - You can paste right back into the same range if you'd just like to convert those formulas to values there
4. Open the Paste Special dialogue either by typing Control<sup>2</sup> + Alt<sup>3</sup> + V. Or:
  1. Click on the File tab, and then the little down-pointing arrowhead under the Paste button
  2. Either choose the type of Special pasting you'd like to do, or clicking on the Paste Special button
5. Select to paste Values (or just type V)

### 2. Transposing Data

Transposing means to flip down from, e.g., going from top to bottom instead to be going from left to right. To do this:

1. Select & copy the range of cells to transpose
2. Move to where you'd like them transposed
3. Access the Paste Special dialogue (e.g. Control + Alt + V)
4. Type T to select transpose, or select that option from the dialogue box.

- More at, e.g., [ablebits.com](http://ablebits.com) or [extendoffice.com](http://extendoffice.com)

## 15.4 Navigation

### 15.4.1 Navigating & Selecting *Within* a Sheet

Table 15.1: Key Commands to Navigate *Within* a Sheet in Excel

Key(s)	Function
Control	Go to the beginning/end of a string (column or row) of cells with values in them
Shift	Select cells
Control + Shift	Select all cells in a string

<sup>2</sup>This is usually the Command key on a Mac.

<sup>3</sup>Alt usually translates to Option on Macs

Key(s)	Function
Home	Beginning of row
Control + Home	Top left of sheet
End	End of row
Control + Spacebar	Select entire column
Shift + Spacebar	Select entire row
Alt + Page Down	Move one screen to the right in a worksheet
Alt + Page Up	Move one screen to the left in a worksheet

### 15.4.2 Navigating *Between* Sheets

Table 15.2: Key Commands to Navigate *Between* Sheets in Excel

Key Combination	Function
Shift + F11	Insert new sheet (to the right)
Control + Page Down	Switch to next sheet to the <b>right</b>
Control + Page Up	Switch to next sheet to the <b>left</b>

More key shortcuts are at:

- [This site with a few of the most common](#) and
- [This MS site](#)

## 15.5 Formulas

Formulas in Excel and Calc allow you to quickly conduct many, useful operations on your data. It's even sometimes easier to use these here than to do them in SPSS, R, or an other stats program. Using formulas in a spreadsheet program helps you get up close and personal with your data before analyzing them in those stats programs, too.

I will cover a few that I use most often in this section. Among the additional resources to consider are:

- [ExcelGPT](#) (aka ExcelBrew), which uses AI to create formulas based on your description of the action(s) you'd like to do

### 15.5.1 General Steps for Entering Formulas (and Common Formulas to Use)

To begin to enter a formula in a cell:

- Type = within a cell to enter a formula
- Alt, M: Go to formula tab
- Common Formulas:

Table 15.3: Common Formulas in Excel

Function	Excel Formula
Sum of range	=sum( <i>cell range</i> )
Minimum value of range	=min( <i>cell range</i> )
Maximum value of range	=max( <i>cell range</i> )
Number of cells <i>with numbers</i> in a range	=count( <i>cell range</i> )
Number of cells <i>with words</i> in a range	=counta( <i>cell range</i> )
Mean of range	=average( <i>cell range</i> )
Standard deviation of range <sup>4</sup>	=stdev.s( <i>cell range</i> )
Generate a random number <sup>5</sup> from 0 – 1	=rand()
Generate a random <i>whole</i> number between a range	=randbetween( <i>lower value,upper value</i> )

### 15.5.2 if

The if function evaluates a cell and returns different values if the contents of that cell do or do not match some criterion.

For example, I could have gender coded in one cell as a 1 if that person identifies as female or a 0 if they identify otherwise<sup>6</sup>. I may then want to have an other cell that actually spells out what that 1 and 0 mean in case I forget. I could use if to do this.

The general format for an if function is:

=if(*Some **statement**,What to output if the statement is **true**,What to output if the statement is **false***)

To continue with that example, let's imagine that the cell with the dummy-code for female is in cell A2 and I want to put the phrase Female or Not Female in cell B2, like this in Excel:

Row/Column	A	B
<b>1</b>	Dummy Variable	Category Label
<b>2</b>	1	Female

To do this, I would type the following into cell B2, i.e., the cell where I want the result to go:

```
=if(A2 = 1, "Female", "Not Female")
```

<sup>4</sup>The .s at the end of it denotes that it is to generate the standard deviation of a sample—not the population. (That's =stdev.p although you'll likely never use it: We rarely have all of the data for a population. That's for qualitative research to say that their participants are that entire population.)

<sup>5</sup>Remember that computers can't do random things. (So, there's no need for a *Voight-Kampff* test; all you need is to detect non-randomness.) So, the values generated by Excel cannot be considered random for, e.g., randomizing within a study. True randomization can be achieved best through old-fashioned ways: throwing dice, pulling pieces of paper from a bag, flipping coins, etc.

<sup>6</sup>This is “dummy coding” gender into a yes/no variable about whether the person identifies as female. If the person identified as female, we code that as a 1; any other response (except missing data) we code as a 0; missing data are coded as missing data. We could also have an other variable that dummy-codes whether they identify as male (there, 1 for male and 0 for anything else.). And yes, this allows us to have times when someone identifies as *both* female and male by letting that person have a 1 for both variables. One of the advantages of dummy coding then is that it allows for multiple responses/categories.

To break down the parts of an if statement doing this:

```
=if(
```

- This first part starts with an equal sign (=) letting Excel know that you'll be entering a function
- The if( lets it know what the function is you'll be using

```
A2 = 1,
```

- This tells Excel where to look to evaluate an argument (we're telling it to look in cell A2)
- It then tells Excel what the formula is to evaluate. Here, we're asking whether the value in cell A2 is equal to 1 or not (= 1)

```
"Female",
```

- This is the value we're asking Excel to return if indeed the value in cell A2 equals 1. If A2 equals one, we want Excel to print out the word Female.
  - Since we're asking Excel to print out a word, we have to put it in quotes.
  - If we'd asked Excel to print out a number, we would *not* put it in quotes<sup>7</sup>.

```
"Not Female")
```

- This is what we want Excel to print out if our formula (A2 = 1) is incorrect, i.e., if the value in cell A2 is *anything* besides a 1.

If we had more cells in that first column—other gender identities for other participants, it could look like the following table—noting that I've added in a column to show what the function would look like right before what Excel would produce in that column:

Row/Column	A	B	
1	Dummy Variable	Function Typed into Cell in Column B	Excel Output in Column B
2	1	=if(A2 = 1, "Female", "Not Female")	Female
3	0	=if(A3 = 1, "Female", "Not Female")	Not Female
4	3	=if(A4 = 1, "Female", "Not Female")	Not Female
5	NA	=if(A5 = 1, "Female", "Not Female")	Not Female
6		=if(A6 = 1, "Female", "Not Female")	Not Female

<sup>7</sup>Unless we instead wanted Excel to treat the output not as a number but as a word, i.e., treating some number as a word and not as a number.

- Note that cell A6 is empty, i.e., row 6 has an empty cell in column A. Excel interprets an empty cell as not meeting the evaluated condition (here that  $A6 = 1$ ).

if statements are very useful. They can evaluate several operations:

Operator	Meaning	Example	Result
=	Equal to	=if(A2 = B2, "TRUE", "FALSE")	Returns TRUE the value in A2 is the <b>same as</b> the value in B2
<>	Not equal to	=if(A2 <> B2, "TRUE", "FALSE")	Returns TRUE the value in A2 is <b>different than</b> the value in B2
>	Greater than	=if(A2 > B2, "TRUE", "FALSE")	Returns TRUE the value in A2 is <b>greater than</b> the value in B2, e.g., if A2 = 2 and B2 = 1
>=	Greater than or equal to	=if(A2 >= B2, "TRUE", "FALSE")	Returns TRUE the value in A2 is <b>greater than or equal to</b> the value in B2, e.g., if A2 = 2 and B2 = 1 <b>or</b> of B2 = 2
<	Less than	=if(A2 < B2, "TRUE", "FALSE")	Returns TRUE the value in A2 is <b>less than</b> the value in B2, e.g., if A2 = 5 and B2 = 10
<=	Less than or equal to	=if(A2 <= B2, "TRUE", "FALSE")	Returns TRUE the value in A2 is <b>less than</b> the value in B2, e.g., if A2 = 5 and B2 = 10 <b>or</b> if B2 = 5

In addition to recoding, you can use it to test if cells are blank (=if(A2 = "", "Blank", "Not Blank")), compute different formulas based on different values (e.g., return the [absolute value](#) if a cell is negative: =if(A2 < 0, abs(A2), A2)), *etc.*

### 15.5.2.1 Nesting if Statements

if statements can be [nested](#). For example, I use the following formula to compute letter grades from a percent grade given in cell A1. I can just drop this formula into, say, cell B1 and it will automatically give me the letter grade using [CUNY's conversion standards](#)<sup>8</sup>:

```
=IF(A1>=97.5,"A+",IF(A1>=92.5,"A",IF(A1>=90,"A-",IF(A1>=87.5,"B+",IF(A1>=82.5,"B",IF(A1>=80,"B-",
↪ IF(A1>=77.5,"C+",IF(A1>=70,"C",IF(A1>=60,"D",IF(A1<60,"F")))))))))))
```

Note that there is an other Excel function, [ifs](#) that makes the syntax for nesting if statements a bit cleaner, but I personally prefer seeing it all spelled out—even if it means having a bunch of parentheses at the end.

The ifs statement for that same coding of percents into letter grades is:

```
=IFS(A1>=97.5,"A+",A1>=92.5,"A",A1>=90,"A-",A1>=87.5,"B+",A1>=82.5,"B",A1>=80,"B-",
↪ A1>=77.5,"C+",A1>=70,"C",A1>=60,"D",A1<60,"F")
```

<sup>8</sup>Notice that the IF statements are read by Excel from left to right, so, if a percent is not  $\geq 97.5$  then Excel goes to the next IF statement to see if it's  $\geq 92.5$ , and if not then goes to the next IF, etc.

### 15.5.3 vlookup

I love vlookup. It's such a powerful way to recode variables based on some criterion. For example, I could recode all males to 0 and females to 1 in just a few steps.

vlookup is a vertical lookup that you use to fill in values down a column. You can use hlookup to look up values to a row (not a column) of data with values also looked up in rows of data. (Apparently xlookup is a new Excel command that knows which way the data are being read and looked up, so can be used instead of both vlookup and hlookup, but I've not gotten it to work.)

- Formula:

`=vlookup(cell to look up, range of cells to find replacement value, column in range for value to return, match mode)`

- Match modes:

Table 15.7: vlookup Formula Match Modes in Excel

Value to Enter	Description
0	Exact match. If none found, returns #N/A. This is the default
-1	Exact match. If none found, return the next <b>smaller</b> item
1	Exact match. If none found, return the next <b>larger</b> item
2	A wildcard match where *, ?, and ~ are “wildcards”, MS’s pathetic attempt at <a href="#">regular expressions</a>

- To “lock” part of a cell reference, add \$ right before it
- E.g.:
  - To keep the reference for replacement values “locked” to cells A1 through B12 (written in the formula as A1:B12) when filling in cells below with that formula,
  - Add a \$ before the 1 and the 12, making it A\$1:B\$12

#### 15.5.3.1 Index and Match

You can get [similar results](#) by instead using the index and match commands. Since I prefer vlookup, for now, I will simply link to this [ExcelJet page](#) that covers them well.

## 15.6 Pivot Tables & Charts

[Pivot tables](#) are a great way to quickly generate descriptive statistics for categories. They are also flexible so that you can look at subgroups or “cross tables” that show, e.g., the stats for two variables that are crossed with each other (like looking at, say, the hip-waist ratios of genders crossed with ethnicities/races).

### 15.6.1 Inserting/Creating a Pivot Table or Chart

#### Table:

- Insert tables using:
  - Alt, N, V
- GUI:
  1. In the Tables group, click on the Insert tab
  2. Click on PivotTable
  3. Click OK
  - [More here](#)

#### Chart:

- Insert charts using:
  - Alt, N, S, C
- GUI:
  1. In the Tables group, click on the Insert tab
  2. Click on PivotChart
  3. Click OK

#### 15.6.1.1 More guides on pivot tables:

- <https://magoosh.com/excel/excel-pivot-chart/>
- <https://blog.hubspot.com/marketing/how-to-create-pivot-table-tutorial-ht>
- <https://www.gurug9.com/pivot-tables-in-excel-beginner-s-guide.html>

## 15.7 Basic Statistics

- Most of the statistical analyses are within the Data tab under the Data Analysis button

### 15.7.1 Correlations

Function	Formula
Return a correlation matrix	<code>=correl(cell range 1, cell range 2)</code>

Creating correlations (and many other descriptive & inferential stats) is also accessible via a GUI in the Analysis ToolPak described under ANOVAs, below.

### 15.7.2 *t*-Test

This formula returns the *p*-value for the *t*-test. It does not return the actual *t*-score.

Formula:

=*t.test(group 1 column range,group 1 column range,tails,type)*

- *Tails:*

Table 15.9: *t*-Test Tails in Excel

Value to Enter	Description
1	One-tailed test I.e., that Group 1 values are larger than Group 2 values
2	Two-tailed test I.e., that Group 1's values are <b>either</b> larger <b>or</b> smaller than Group 2's values

- *Type:*

Table 15.10: *t*-Test Types in Excel

Value to Enter	Description
1	Paired <i>t</i> -test I.e., that the values in a given row are from the same participant. E.g., Group 1 is a person's pretest score and Group 2 is that same person's posttest score
2	Unpaired <i>t</i> -test assuming homoscedasticity I.e., that the variance in the populations from which Group 1 is sample from is the same as the variance in the population from which Group 2 is sampled
3	Unpaired <i>t</i> -test assuming heteroscedasticity I.e., that the variances in the populations from which Groups 1 and 2 are sampled are different

N.b., OLS tests (e.g., *t*-tests & ANOVAs) are rather robust to heteroscedasticity, so choosing 2 for unpaired *t*-tests is generally fine. If variances are very different, then standardizing the variables usually suffices to address the issue.

#### 15.7.2.1 Steps to Conduct a *t*-Test on the China Posttest Data

As an example for computing the *p*-value for a *t*-test in Excel, let us use the posttest data from the study of elementary students in China.

1. Open the CFL\_Posttest\_Data.csv file in Excel
2. Sort all of the data by Population. To do this:
  1. Type Control + Home to go to the top-left cell
  2. Shift + Control + End to select all of the data

3. Type Alt (Option on a Mac), then D, and then S to open the Sort dialogue box (it's not easy to see how to get the combination, but it's one I use often enough to have just memorized, If it helps to know, that's D for the Data menu and then S for Sort.
  4. Sort by Population in A to Z order
3. Now go to somewhere outside of the data to create the formula. Me, I created a new sheet and went to that instead of doing things on the sheet with the data—don't want to accidentally overwrite data and not realize it. Believe me.
  4. In essence, the `t.test` formula has you type `=t.test()`<sup>9</sup> and then to follow that with a selection of the first set of data and then follow that with a selection of the second set of data. Plus some other, almost-random stuff at the end of the command.
    1. You can do all of this by typing `=t.test()` into a cell and then going back to the sheet with all of the data left-mouse-clicking in the first cell with `toqa.pro` data (cell M2), and then holding the mouse button down while you scroll down until you see in the Population column that it's changed from Migrant to Non-Migrant to select the first group of data. Then repeating that for selecting the non-Migrant data in below that in the same column.
    2. Or you can paste this into a cell:  
`=T.TEST(CFL_Posttest_Data!M2:M849,CFL_Posttest_Data!M850:M1130,2,2)`  
 What that does is say:
      1. The first set of data are in the `CFL_Posttest_Data` sheet in cells M2 through M849; these are the Migrant students' raw prosociality scores
      2. The second set of data are also in the `CFL_Posttest_Data` sheet, these in cells M850 down through cell M1130; these are the Non-Migrant students' raw prosociality scores
      3. That first 2 after indicating the second set of data tells Excel that it's a two-tailed test (meaning we're testing whether the Migrant scores are higher or lower than the non-Migrant scores; a one-tailed test would test only if they were higher but not lower; one-tailed tests are more powerful but only answer that one question—as is common in stats, there's a trade off between power and precision)
      4. That second 2 tells Excel it's an unpaired *t*-test (meaning the people in one group are not the same people as in the second group; they could be the same people, e.g., if we were following the same people in a pre-post design). It also tells Excel to assume that the variance of prosociality scores is the same in both groups; again, this is an assumption that's (a) usually wrong to some degree and (b) O.K. to make as long as they're not that different.
    3. After doing all this mountain of work, you should be gifted with a rather anticlimactic result of a single number appearing in that cell. That modest number should be 0.196848 (depending on how many digits you see). That is the *p*-value for the *t*-test; if this number were less than, say, .05, we would say that there is a significant difference between the two groups. However, it is not, so we would not say there is a significant difference.

### 15.7.3 ANOVAs

#### 15.7.3.1 Installing the Anlaysis Add-in

The functionality to conduct most of the stats that Excel can are not loaded by default. (These are, however, available by default in [LO Calc](#)). Fortunately, they are (at least so far) available through an easy installation. To do this:

---

<sup>9</sup>Typing the left parenthesis let you then fill in the values with mouse selections and/or further typing

1. Click on File
2. In the menu that opens, click on Options then Add-ins
3. Select the Analysis ToolPak option near the top
4. At bottom of dialogue box that opens, under Manage, select Excel Add-ins and click Go
5. Select Analysis ToolPak (and whichever other ones you want)
  - (VBA is for using MS's visual basic functionality—a holdover from when Windows & Office were still figuring themselves out)
6. An Analysis tab will now appear in under the Data tab

### 15.7.3.2 Conducting an ANOVA

E.g., a “Single Factor” ANOVA, which is otherwise known (confusingly) as a one-way ANOVA. This is simply an ANOVA that has only one predictor (or independent variable, IV) and one outcomes variable (or dependent variable, DV).

1. Under the Data tab and in the Analysis group, click on the Data Analysis button
2. In the dialogue box that opens, click on Anova: Single factor
3. Click on the icon next to the Input Range field to minimize the next dialogue box that opens
4. Select the range of cells (here, two columns, one for the IV & one for the DV), and then click on the icon to maximize that dialogue box
5. Choose whether data are grouped by columns or rows—most likely by columns, with one column for each variable
6. Choose where the results will be posted; I tend to choose New Worksheet Ply: which I give an apposite name to, e.g., “ANOVA Source Table”
7. Click OK





## Chapter 16

# Introduction to Power and Sample Size Estimation Using Either G\*Power or R

### 16.1 The Relationship Between $\alpha$ , Power, Effect Size, and Sample Size

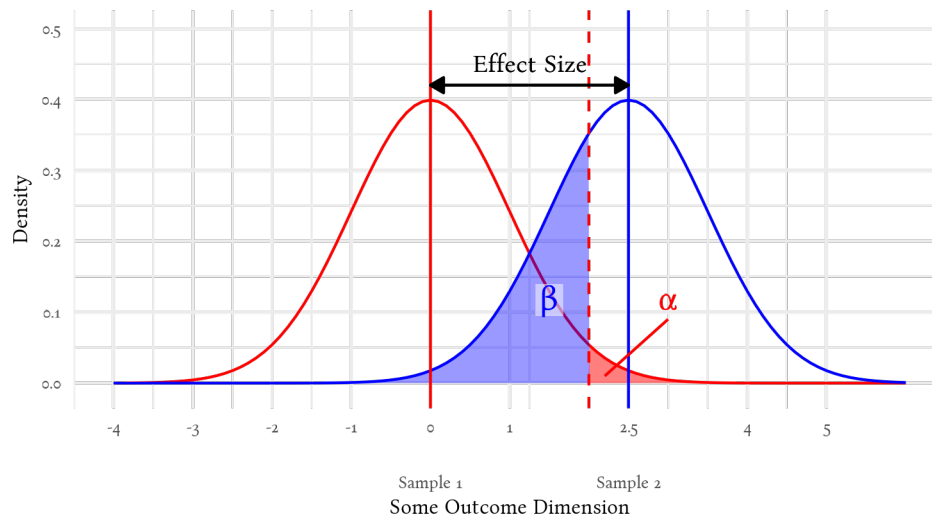
Power is one of four, inter-related values used (implicitly or explicitly) in hypothesis testing:

1.  $\alpha$ , the probability of a false positive—seeing an effect that isn't there. (Also called a Type 1 error.)
2.  $\beta$ , the probability of a false negative—missing a real effect. (Also called a Type 2 error.)

- Note that power is  $1 - \beta$

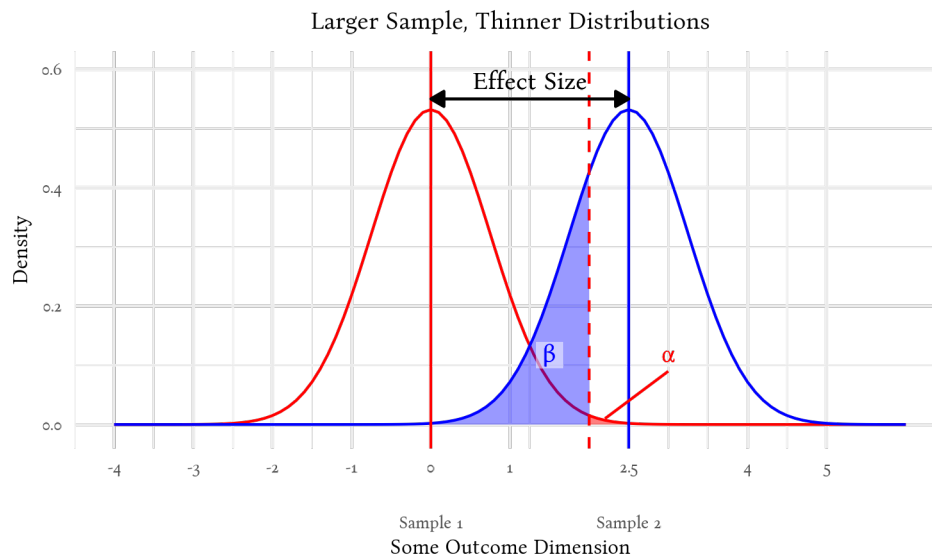
3. **Effect size**, the magnitude of a measured effect. More about this is at Chapter 6
4. **N**, the sample size.

Changes in any of these four values affects the chances of obtaining a significant effect:



Sample size doesn't appear directly in this figure, but it affects the width of the distributions—when those distributions represent uncertainty around an estimate (such as estimates of a population mean<sup>1</sup>). We are using the sample values to estimate these population values, so these distributions represent the probabilities of what the population values are given the sample values we found. The vertical red and blue lines denote the sample values we got for each group, and the curves represent the probabilities of where the actual population values are expected to be. The larger our sample, the more confident we are that our sample's values are darned close to the population values—and so the distributions become thinner.

As the samples grow—and the distribution of estimates of the population means become thinner—the chances of false positives ( $\alpha$ ) and false negatives ( $\beta$ ) both become smaller:



<sup>1</sup>Although we often think of figures like this showing differences in estimates of means, these could instead be estimating, e.g., population proportions (whether, e.g., the proportion of members of two populations have different mortality rates for a given disorder). Since we tend to think about this in terms of means, though, let's just stick with that.

Generally, if we know any three of those values— $\alpha$ ,  $\beta$ , effect size, or  $N$ —we can compute the fourth<sup>2</sup>. This most often means that we can estimate the sample size ( $N$ ) that we would need to detect a given effect size, while assuming particular values for  $\alpha$  and  $\beta$ .

## 16.2 Using G\*Power or R to Estimate a Priori Sample Size Estimates

G\*Power (and R) are perhaps the best, current, one-stop applications<sup>3</sup> to estimate sample sizes needed to expect significance of many, common analyses.

### 16.3 G\*Power

G\*Power is a free (as in “free beer”<sup>4</sup>) software follows the “Unix” philosophy to do one thing and do it well. What it does well is estimate how large a sample one will need to be for various analyses. As the name implies, it’s in fact designed to conduct analyses related to statistical power, but it most often used to computes sample sizes well (power is only occasionally worth computing anyway).

#### 16.3.1 Installing G\*Power

Finally, software that’s easy to install:

1. Click the [Download](#) button on the right of the [G\\*Power](#) site
2. Download the latest version for either Mac or Windows<sup>4</sup>
3. The downloaded file is zipped, so extract it and install.

#### 16.3.2 Citing G\*Power

The creators of G\*Power [request](#) that one uses one or both of the following citations when using it:

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). [G\\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences](#). *Behavior Research Methods*, 39, 175–191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). [Statistical power analyses using G\\*Power 3.1: Tests for correlation and regression analyses](#). *Behavior Research Methods*, 41, 1149–1160.

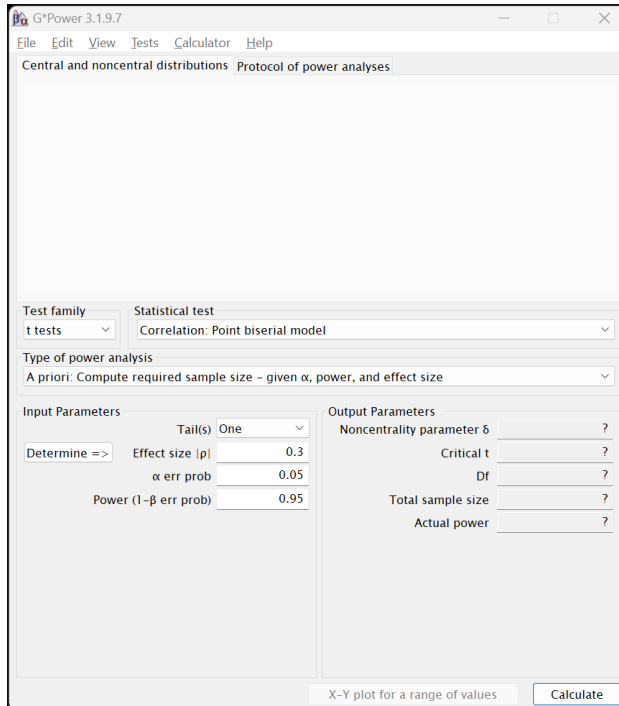
<sup>2</sup>Some analyses are not this straight forward, though, for various reasons. Among the reasons why it isn’t always this straight forward is because the  $N$  for the sample is divided out unevenly among the effects being tested—especially for nested effects; we will address this in one of the cases you are more likely to encounter (it’s different types of ANOVAs), but first let’s orient ourselves to G\*Power and conduct some more straight-forward sample size estimates.

<sup>3</sup>Well, to be pedantic, R may be considered [software](#).

<sup>4</sup>Alas, there is no GNU/Linux version.

### 16.3.3 Orientation to G\*Power

G\*Power has a rather simple interface. Below the menu bar at the top (and a window that will fill with power curves for your estimation), it presents a drop-down menu for Test family and Statistical test; most of the options below these drop-down menus will change depending on the options chosen there.



Usually, the first step to use G\*Power is to select Test family and then the Statistical test. After that, one fills in the values in the fields below in the Input Parameters section.

#### 16.3.3.1 Type of power analysis

The drop-down menu under immediately under Test family and Statistical test allows one to compute either sample size or power,  $\alpha$ , or effect size. The only option we will use is A priori: Compute required sample size - given  $\alpha$ , power, and effect size. However, it's worth briefly discussing the other options.

The Type of power analysis offers more options than the A priori: Compute required sample size - given  $\alpha$ , power, and effect size. Most of the options pertain to computing the other variables in the  $\alpha$ , power, effect size, sample size group. There is one worth explaining, however: Compromise : Compute implied  $\alpha$  & power - given  $\beta/\alpha$  ratio, sample size, and effect size

#### A priori: Compute required sample size and a Note on Post hoc Power Analyses

Computing effect size can be informative. Computing required  $\alpha$  may be interesting if rarely practical. Computing post hoc “achieved power,” however, is rarely either [hoenig2001]. Post hoc power analyses depend very heavily on the particulars of a given set of data; i.e., they do not generalize—neither to other samples nor even to the actual power of the test (Yuan & Maxwell, 2005; Zhang et al., 2019).

Post hoc power analyses are also based on a tenuous interpretation of power: Achieved power is computed assuming that there is an effect, but hypothesis tests actually assume that there is no effect. Remember that hypothesis tests actually test the probability of finding the results we did *if the null hypothesis is true*. If we conclude that there is a significant effect, we are saying that there is insufficient evidence to conclude that the null is true. We are thus also stating that there is insufficient evidence that a post hoc power analysis is justified.

#### **Compromise: Compute implied $\alpha$ & power**

The Compromise: Compute implied  $\alpha$  & power - given  $\beta/\alpha$  ratio, sample size, and effect size option is an interesting one—even if it's rarely useful. We begin with a sample size—typically the largest sample we know we would be able to attain within practical constraints—and an effect size that we either know or expect to have. Within those real constraints we can explore what levels of  $\alpha$  and  $\beta$  we could achieve. We could see—for example—how badly power would be affected if we try to maintain a significance level of .05, or what level of significance we would have if we tried to keep the chances of both types of error the same.

It is an idea that harkens back to the original intents of those (like Fisher) who originally thought up the idea of a significance test. The significance level wasn't always frozen at .05; the original idea was to use whatever level one felt was appropriate—be it .05, .01, .10, .25 or any level that conveyed how important one felt false positives were in a particular situation.

But messing with  $\alpha$  has become verboten. A wall doing its best to keep out *p*-hackers and other all-too-human threats to the integrity of science. There is now little practical point to play with the proportion of false positives. It may satisfy an idle curiosity to see how strongly it is affected versus false negatives, but it isn't going to get you any closer to published or funded. Let's instead explore what does.

#### **16.3.3.2 A Note About the Default Level of $\beta$ in G\*Power**

It is common in the health and social sciences to also assume power is .80, i.e., that we have an 80% of detecting a real effect. Otherwise said, we assume we have a 20% to *miss* detecting a real effect. The convention is to prefer to miss seeing something important over mistakenly thinking we found something; we accept more false negatives ( $\beta = .20$ ) than false positives ( $\alpha = .05$ ). This convention is not rigid, however, and we certainly can and should change those values—such as making  $\alpha$  and  $\beta$  the same—if justified.

However, the default value for  $\beta$  in G\*Power is .05, thus setting the chance of a false negative ( $\beta$ ) the same as the chance of a false positive ( $\alpha$ ). G\*Power is used in many areas of science, and that .80 convention for power isn't followed everywhere. For example, pharmaceutical researchers often use that higher value for power ( $1 - \beta = .9$ ) so they have a better chance to detect real—even rare—side effects in clinical trials.

#### **16.3.4 Estimating Required Sample Sizes**

For all of these exercises, we will be estimating how large a sample we would expect to need in various analyses and assuming common standards for  $\alpha$  and  $\beta$ . Before we do, however, I want to point out that the key words in that sentence are “estimating” and “expect”: Sample size estimation is *not* an exact science and the actual study we conduct will surely have actual rates of significance different from what we expect. After all, if we already knew what we would find, we probably wouldn't be conducting the experiment.

It is also worth pointing out that we pretty often actually need a larger sample than we estimate. There is a tendency to be overly optimistic about our abilities to realize certain effects or overcome real-world challenges for recruiting a conscientious group of well-delimited participants. Nonetheless, our a priori estimates can at least put us in the general vicinity of where we need to be.

### 16.3.4.1 Correlations

Let us begin with estimating samples for correlations. Although it's not evident from most stat programs, we use different sorts of tests for different sorts of correlations. Therefore, tests of correlations are under different Test family options in G\*Power.

#### Pearson's Product-Moment Correlation ( $r$ )

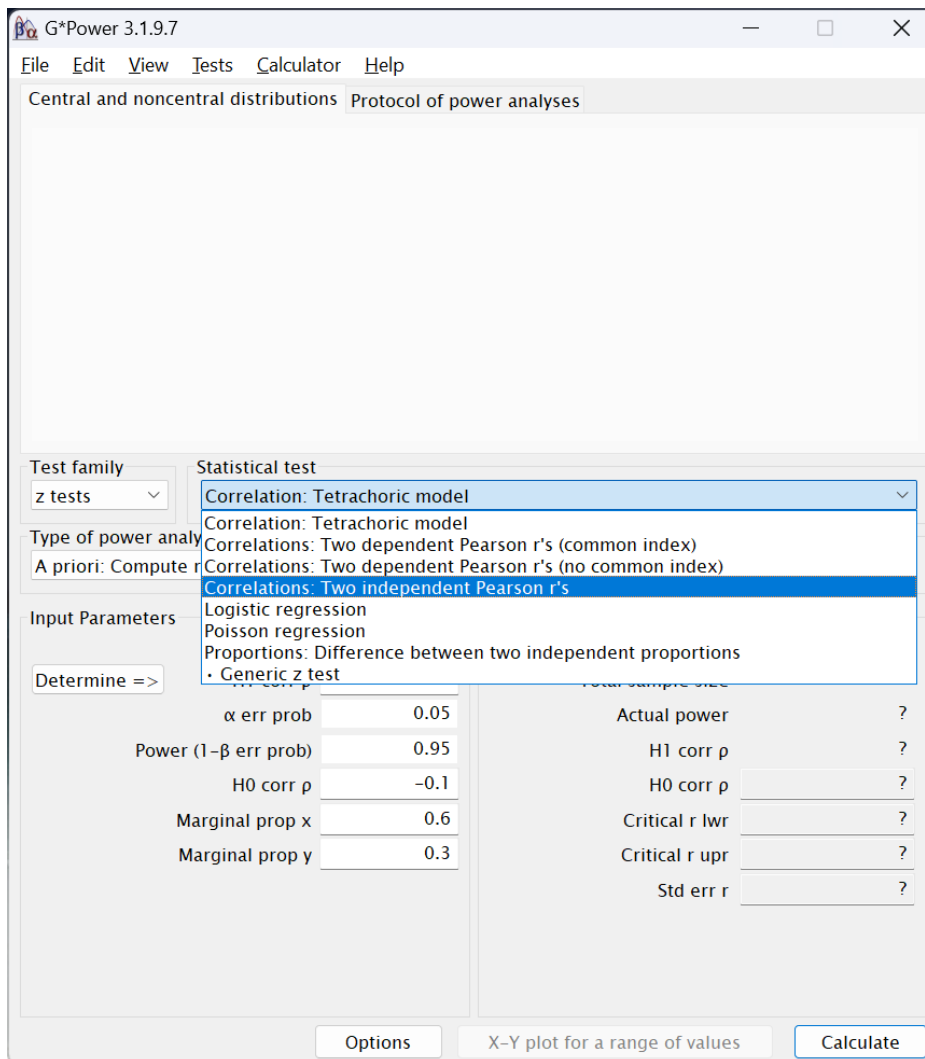
We begin with the most commonly-used type of correlation, that which is most formally called Pearson's product-moment correlation coefficient. This is the one symbolized with a simple  $r$  that measures the degree of association between two continuous variables. This sample size estimate is for tests of whether the correlations differ between two groups.

For example, if Hepatitis B/C is more strongly correlated with incidents of liver cancer in [males than in non-males](#). Or whether [the correlation between income and longevity is different](#) among Blacks versus that correlation among whites. Let's use that latter example and assume that we expect that the correlation will be .5 among Blacks but only .3 among whites (or vice versa). We're thus interested in seeing how large a sample we would need to reliably detect a significant difference between these correlation coefficients.

1. It is found under the z tests option in the Test family menu.
2. After selecting that option, under Statistical test, choose Correlations: Two independent Pearsons's  $r$ 's<sup>5</sup>:

---

<sup>5</sup>Even though using an apostrophe in  $r$ 's is incorrect: It's plural, not possessive, so should instead be  $r$ s. Just one of my many grammatical pet peeves.

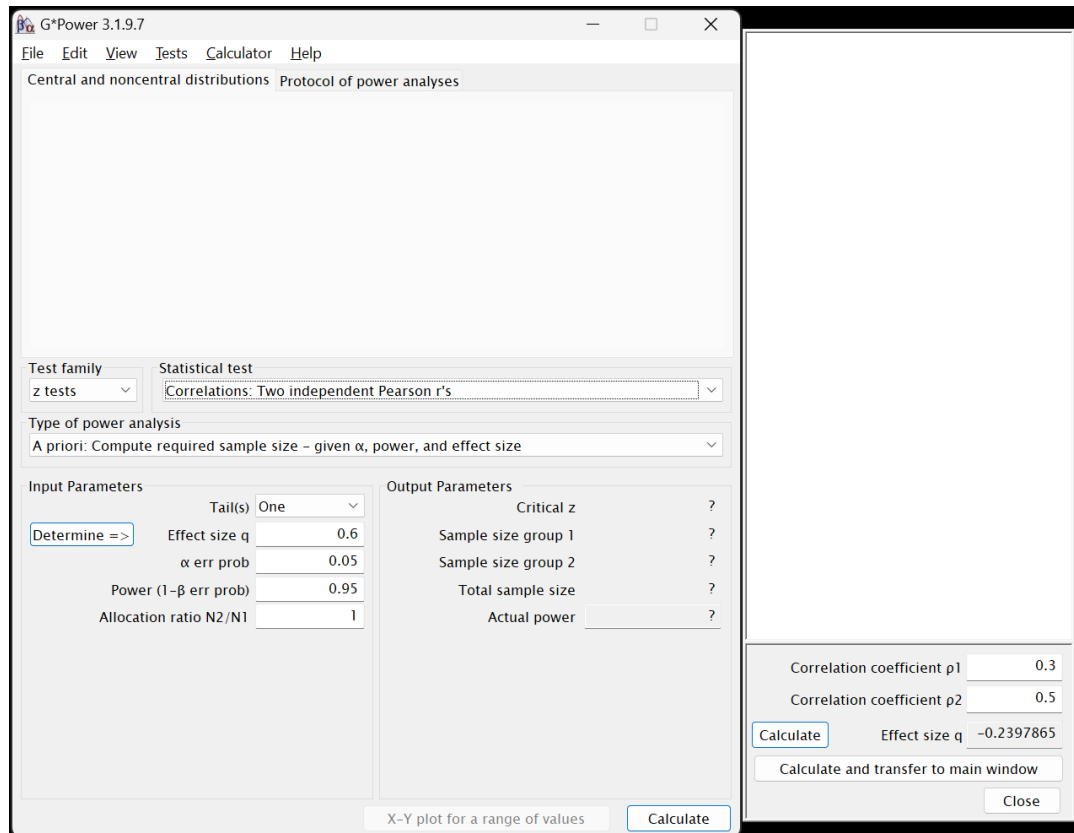


3. As we will for all of these exercises, under Type of power analysis, choose A priori: Compute required sample size - given  $\alpha$ , power, and effect size
4. In the Input Parameters section, let's first choose One under Tails. (We will next see how things change when we choose two.)
5. Handling effect size

**1. By computing the difference between two anticipated correlations.**

1. The effect size statistic for a difference between correlations is Cohen's  $q$  (Cohen, 1988, p. 109), as noted in the next row, Effect size  $q$ . Most effect size statistics are easy to compute, but unfortunately  $q$  is not<sup>6</sup>. Fortunately, G\*Power can easily compute it for us:
2. Click on the Determine => button to the left of Effect size  $q$ . A new window will open to the right of the main window:

<sup>6</sup>It's this monstrosity:  $q = \frac{1}{2} \log_e \frac{1+r_2}{1-r_2} - \frac{1}{2} \log_e \frac{1+r_1}{1-r_1}$  for correlations  $r_1$  and  $r_2$ .



3. By sheer coincidence, the default values given by G\*Power for the first (Correlation coefficient  $\rho_1$ ) and second (Correlation coefficient  $\rho_2$ ) correlations are the ones we want<sup>7</sup>, so leave them as they are.
4. Still inside the side window, click on the Calculate and transfer to main window button. The computed value for  $q$  will be populated into both the side window's and the main window's Effect size  $q$  fields.
5. Click Close in that side window.  
Note that most power analyses in G\*Power will have a Determine => button, but they will do different things for different tests. Often, it is used to convert one effect size measure to an other one that is more appropriate for the given test (unlike we did here).

## 2. By using Cohen's (1988) recommendations.

1. Back in the Input Parameters section of the main window, mouse over the field where we will enter our values:

<sup>7</sup>Changing the order of them so that Correlation coefficient  $\rho_1$  is .5 and Correlation coefficient  $\rho_2$  is .3 will make the effect size measure,  $q$ , into a positive value, but that doesn't matter: The positive or negative sign simply indicates which correlation is larger than which. Making it positive will also move the  $H_1$  distribution to the right of the  $H_0$  one in the figure at the top.

The bubble that appears lists the values that Cohen (1988) recommends for—in this case—Pearson’s  $r$ s. A bit more about his recommendations for these is given on [page 129](#) of his book. Looking at those suggested levels, we see that the effect size for a difference between  $r_1 = .3$  and  $r_2 = .5$  is a bit less than a “medium”-sized effect.

2. Although Cohen’s recommendations are viewed by most as more canonical than Cohen intended, it is still useful practice to use them as guides for anticipating effects, especially if we don’t know ahead of time what size of effects we can expect. Or here, if we didn’t have any reason to expect certain correlations for each group beforehand.

In such cases, researchers often assume *a priori* that they will achieve a “medium” effect. (Although I tend to recommend assuming something half way between a “small” and “medium.”) G\*Power also uses a “medium” effect as well for the default value, so feel free to use that here instead of the  $\approx |0.23|$  that we computed from *a priori* expected correlations.

It is quite worth noting, however, that it is greatly preferable to instead use prior research—even if only tangentially related—to estimate what levels of effects one should expect.

- **For Effect size  $q$ , please enter either  $\sim 0.23$  or  $.3$ .** I will continue assuming that you entered in the latter option.

6. The  $\alpha$ -error prob option can be left at the 0.05 default both now and probably always.
7. The default value in G\*Power for Power ( $1 - \beta$  err prob) is 0.95. Since  $1 - .05 = .95$ , that is implying that the probability of a false negative ( $\beta$ ) should be equal to the probability of a false positive ( $\alpha$ ). I commend the creators of G\*Power<sup>8</sup> for advocating for this parity. However, convention is to deprecate power relative to  $\alpha$  and set  $1 - \beta$  to  $.8$  (thus setting  $\beta = .2$ , four times larger than  $\alpha$ ).

- **tl;dr:** Enter  $.8$  into the Power ( $1 - \beta$  err prob) field

8. The Allocation ratio  $N2/N1$  is asking if the size of the two groups (i.e., the groups whose different correlations we’re testing, e.g., if Blacks versus whites have a different correlation between income and longevity) is the same.

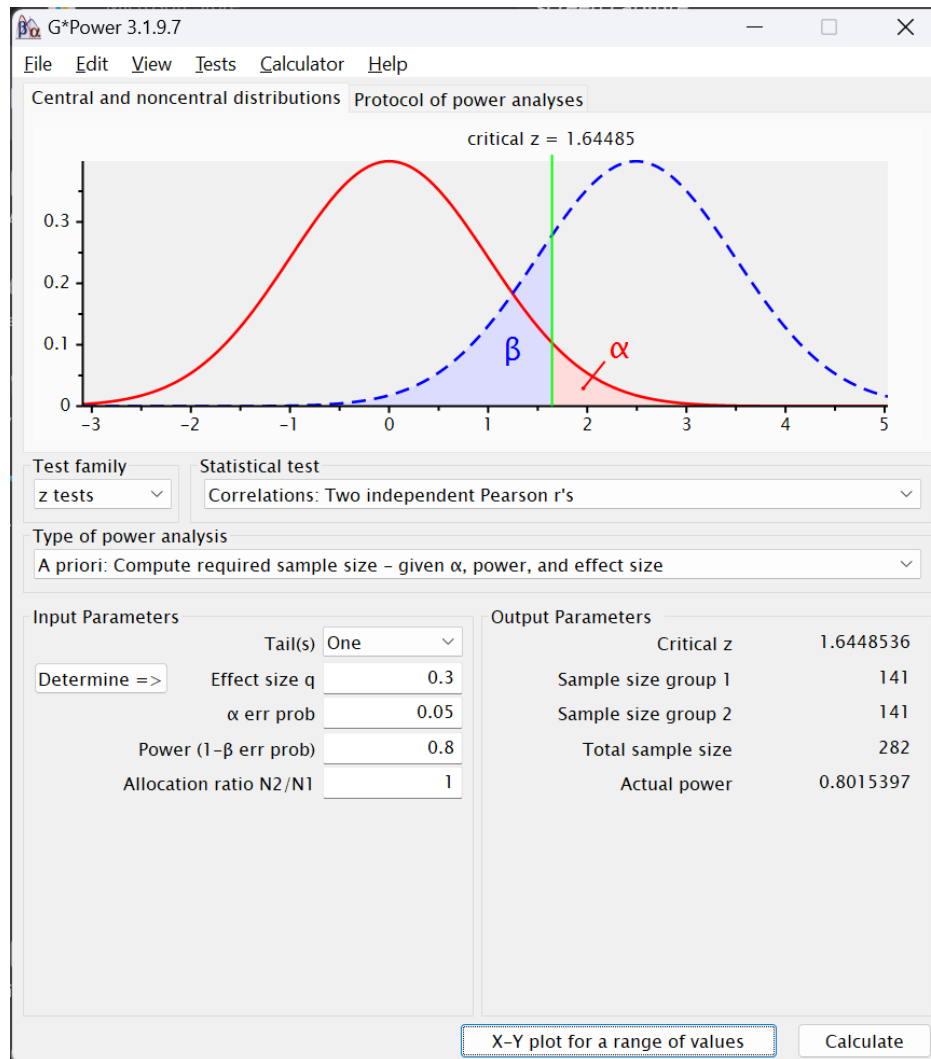
- If we assume that there will be equal numbers in both groups, then enter 1.
- If we assume, e.g., that we will have twice as many whites as Blacks, then enter 2 (or 5 to change which group is considered which race).

<sup>8</sup>G\*Power was created and is maintained by Edgar Erdfelder, Franz Faul, and their colleagues at Heinrich-Heine-Universität Düsseldorf.

- Of course, any other ratio can be used, and different ratios tried to obtain tolerance ranges.

9. Click on the Calculate button in the lower right.

Assuming you used these values—including 1 for Allocation ratio  $N_2/N_1$ —your output should look like this:

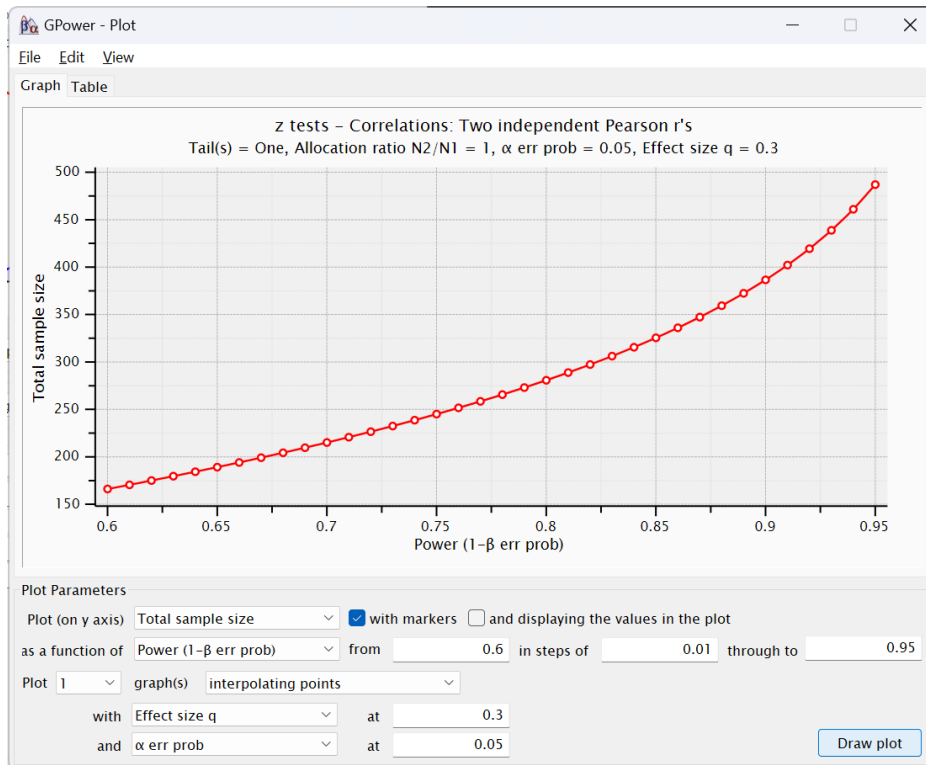


- Under the Output Parameters section, we can see that the Critical z value is  $\approx 1.64$ . This is also presented in the figure near the top of the G\*Power window. This is simply the value of the test statistic (here z) computed and used to test the significance of the difference between correlation coefficients.
- Of greater interest are the next three rows.
  - Sample size group 1 and Sample size group 2 present the required samples sizes for the two groups. Note that these would be different if we had chosen a value other than 1 for the Allocation ratio  $N_2/N_1$ .
  - Total sample size is just the sum of those two samples sizes. We would thus need a total  $N$  of 282 to detect a difference between the two correlations (or 438 if you used  $-.23$  for the effect size).

- The Actual power field simply presents what the power is for these two sample sizes. It differs slightly from what we entered into the Power ( $1 - \beta$  err prob) field due to rounding error from the sample sizes needing to be whole numbers.

If we change the tab at the top from Central and noncentral distribution to Protocol of power analysis, instead of the two curves at the top, we see both the values we entered to estimate the sample size and the output. This alone is of little use, but of slightly more use is that, under this tab, we can select File > Save Protocol and save these input and output values as a text file. (Under the Central and noncentral distribution tab, we can instead save that image of the two distributions.)

Click now on the X-Y plot for a range of values button next to the Calculate button. The following new window will appear:



This figure presents what the power would expected to be were we to use different *total* sample sizes. You will see that the *y*-axis (Total sample size) is about at 280 (or 440 if you used  $q = -.23$ ) when the *x*-axis is at 0.8. Had we used G\*Power's default of .95 for power, the estimated total sample size would have been a little less than 500.

The fields below this figure present the values we entered in the other window as well as options for changing the figure, including changing the range of power values presented and the "steps" between each dot in the figure. The Table tab at the top lets us look at these values as a table instead of a figure for more precision:

I think this X-Y plot for a range of values window is under appreciated. We can change values in this figure (and the corresponding table) to see not only the estimated sample size needed, but how smallish changes to the values—especially to power—would change with different assumptions.

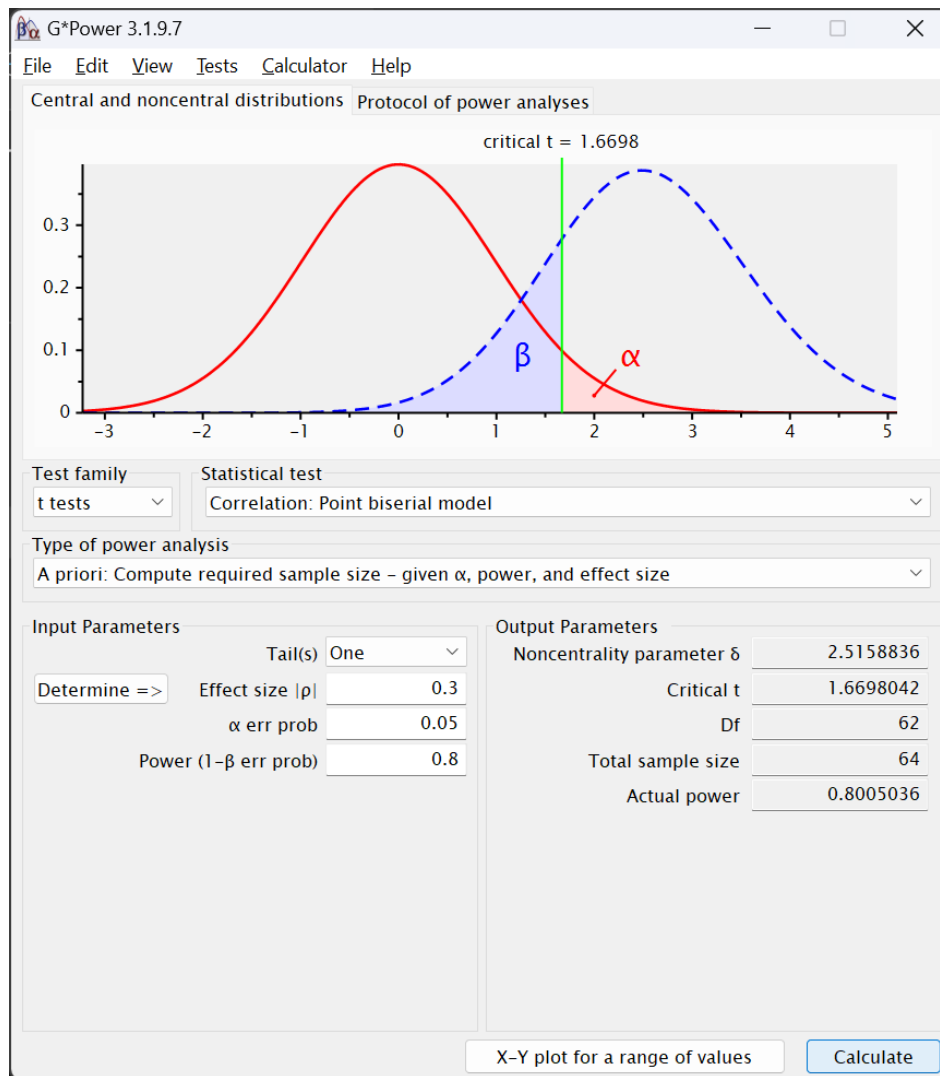
Please return to G\*Power's main window and change the Tails to Two and recalculate the sample size estimation. Under the Protocol of power analysis tab, you can save that output as, e.g., corr\_2-tailed.rtf.

More about this particular analysis is on [page 65](#) of the *G\*Power Manual*.

**Point Biserial Correlation ( $r_{pb}$ )**

Point biserial correlations measure the association between a dichotomous variable and a continuous variable (e.g., the association between pregnancy and blood pressure). In significance tests of point biserial correlations, we're seeing how closely matched some continuous outcome score is for two groups. Two nominal groups and a continuous outcome, that sounds like a  $t$ -test.

1. Under Test family, choose  $t$  tests.
2. Under Statistical test choose Correlation: Point biserial model
3. As always, under Type of power analysis, choose A priori: Compute required sample size - given  $\alpha$ , power, and effect size
4. Since we usually use two-tailed significance tests, please change Tails to Two. (We will next see how things change when we choose two.)
5. The effect size statistic has changed to Effect size  $|p|$  since a different measure is used to compute it for  $r_{pb}$  (and  $t$ -tests in general). Mousing over that field, though, shows that the same values pertain for "small," "medium," and "large" effects. (This isn't always the case.) Again, please enter .3 for a "medium" effect.
6. The only other options here are for  $\alpha$ -error prob and Power ( $1 - \beta$  err prob) for which we will use .05 and .8, respectively.
7. Click on the Calculate button in the lower right. The following should appear:



The Output Parameters area contains somewhat different values. Instead of  $z$ , the test statistic is the Noncentrality parameter  $\delta$  along with Critical  $t$ , which is tested—in this particular case—against a  $Df$  of 62. That is two  $df$  less than the Total sample size estimated to be needed to find an effect under these conditions.

#### 16.3.4.2 $t$ -Tests

Along with a few other tests (like for  $r_{pb}$ ), all  $t$ -tests are under Test family  $> t$  tests. Two of the most common  $t$ -tests are:

- Means: Differences between two dependent means (matched groups)
  - This is used when the values used to compute the two means we're testing come from the same participants, e.g., when we have pretest and posttest measures for the same patients.
- Means: Differences between two independent means (two groups)

- This is used whenever the means come from different participants. This is more often the case—and would apply even if we had pretest and posttest measures but from *different* patients (e.g., if we measured HCAHPS satisfaction scores from an outpatient unit before and after an intervention).

1. Let us use the latter of those two, Means: Differences between two independent means (two groups)
2. The options for Input Parameters is similar to what we were presented with for Pearson correlations (in Section 16.3.4.1). However, when we mouse over the Effect size  $d$  field, we see different recommendations for sizes:

The screenshot shows the G\*Power software interface. The 'Test family' is set to 't tests' and the 'Statistical test' is 'Means: Difference between two'. The 'Type of power analysis' is 'A priori: Given alpha, power'. The 'Input Parameters' section shows 'Effect size d' set to 0.5, 'alpha err prob' set to 0.05, 'Power (1-beta err prob)' set to 0.95, and 'Allocation ratio N2/N1' set to 1. A tooltip for 'Effect size d' lists conventions: d = .20 - small, d = .50 - medium, and d = .80 - large.

- Cohen's  $d$ , the measure of effect size now used, is computed differently than his  $q$  used to measure the effect size of difference between correlation coefficients. As I describe in Chapter 6, this is simply the standardized difference between the two means; it's also one of the most commonly-used measures of effect size.
  - Let's stick with the default given, a "medium" effect of 0.5.
3.  $\alpha$  err prob can of course stay at 0.05
  4. But please change Power ( $1-\beta$  prob) to .8
  5. We could again stipulate a different ratio for the number of participants in the two groups, but let's again assume we will have equal numbers and enter 1 in Allocation ratio  $N2/N1$

Like with  $r_{pb}$ , in the Output Parameters section, we have Noncentrality parameter  $\delta$ , Critical  $t$ , and  $Df$ . However, we now have two groups, so we instead have Sample size group 1, Sample size group 2, and Total sample size. That last field reports expecting to need 102 total participants, so the  $Df$  for the Critical  $t$  is 100.

### 16.3.4.3 F-Tests

$F$ -tests are primarily (nearly only) used to test effects in ANOVAs and their ilk (ANCOVAs, MANOVAs, etc.). These include the most complex sample size analyses available within G\*Power, and the most complex ones researchers typically try to do<sup>9</sup> The reason for this is because the variance associated with effects differ depending on a couple of parameters, such as the number of variables and whether those variables are "fixed" or "random" effects<sup>10</sup>.

<sup>9</sup>I give resources in the Section 16.5 section about estimating sample sizes for factor analyses, etc. These are important...but much less straight forward.

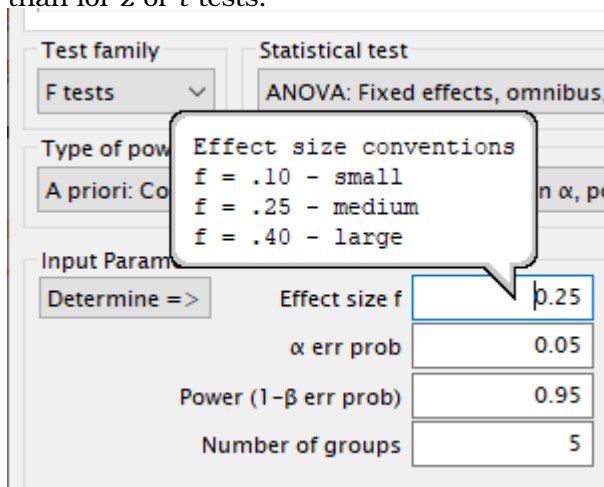
<sup>10</sup>"Fixed" factors are those for which all possible levels are present in the data, e.g., if data for both those who survive and die from an illness are present. "Random" factors are those for which a random subset of all possible levels of a

**ANOVA: Fixed effects, omnibus, one-way**

A one-way ANOVA is simplest form of ANOVA: It contains only one input variable<sup>11</sup> (and one output variable). This one-way ANOVA is really just a *t*-test that is used when the input variable has more than two levels to it. We would use a *t*-test to test the difference (in some continuous outcome) between two groups, say between those diagnosed or not diagnosed with a certain cancer. If there are more than two groups—if, e.g., we were instead looking at the stage of the cancer—we would use a one-way ANOVA.

Sample size estimates for one-way ANOVAs thus closely resemble those for *t*-tests. The difference is that, for one-way ANOVAs, we must indicate how many levels the input variable has. To conduct sample size estimates for them:

1. Under Test family, choose F tests
2. Under Statistical test, choose ANOVA: Fixed effects, omnibus, one-way
3. Under Type of power analysis choose A priori: Compute required sample size - given  $\alpha$ , power, and effect size
4. In the Input Parameters section, when you mouse over the Effect size *f* field, you will once again see that the values Cohen (1988) suggests for “small” through “large” effects are different than for *z* or *t* tests:



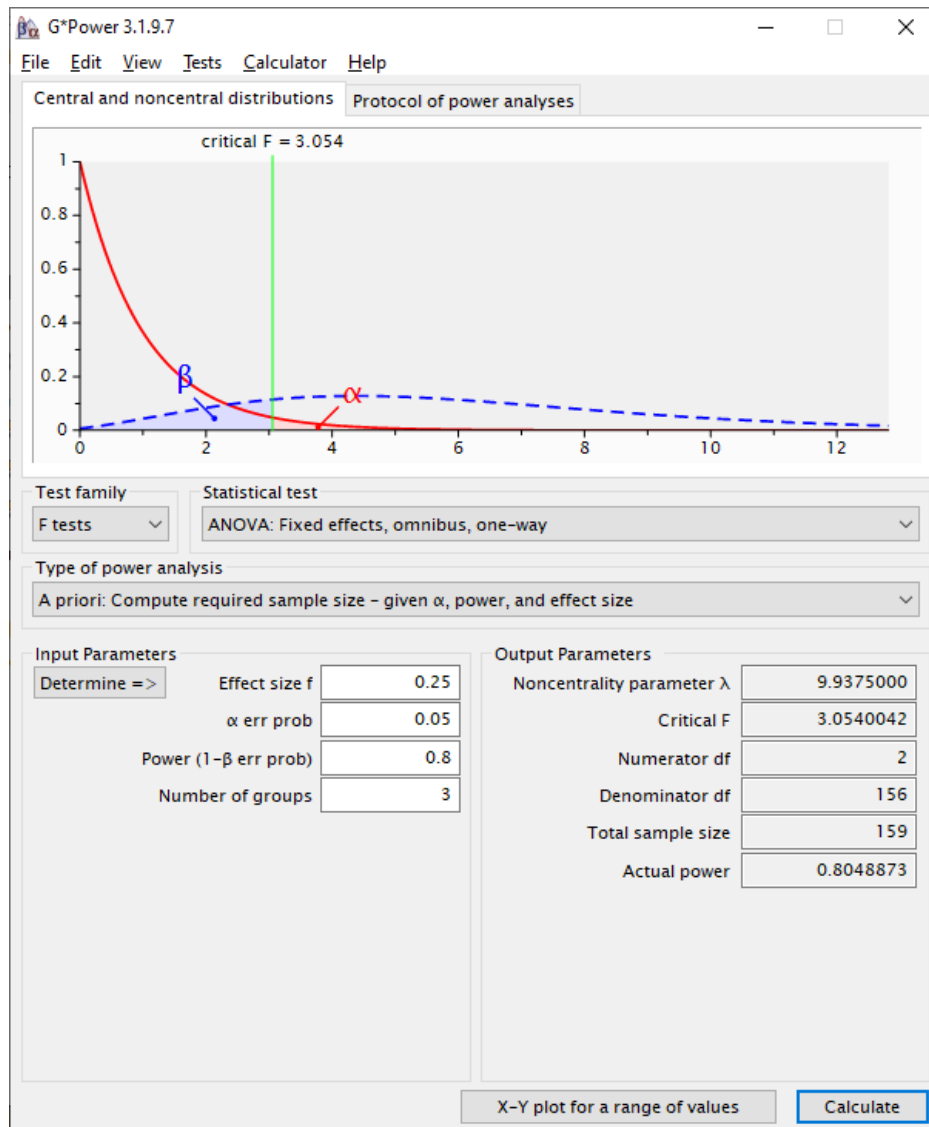
As noted in Chapter 6, Cohen's *f* denotes the effect of a variable after partialing out the effects of other variables. In a one-way ANOVA, there are no other effects—no other variables—but the criteria for effect sizes is still based on this other standard.

- Let's again leave this value at the default—a “medium” effect of 0.25
5. We can also again leave  $\alpha$ -error prob to be 0.05 but change Power ( $1 - \beta$  err prob) to be .8.
  6. The Number of groups field is slightly misleading. This is in fact asking for the number of levels of the input variable. So, e.g., if we are studying the effects of cancer stage and using the TNM staging system, we would have three groups—one for each of the three stages.
- Going with that, please enter 3 in the Number of groups field.

variable are present; these are typically continuous variables (such as height). The ways that error is estimated for these differ.

<sup>11</sup>You may well have learned to call this the “independent variable” (IV). This is not wrong—it's just not always correct. An IV is the term used to indicate a variable that researchers manipulate or measure in a true experimental design to observe their effects on some outcome—that which (in that context) is called a dependent variable and that I'm calling by the more general term of “output variable.” The input variable in an ANOVA is indeed often an IV. However, there are certainly times when we wish to test the effects of a variable that isn't an IV, for example if we're working with secondary data in which no variables are forcibly IVs or DVs.

7. Clicking Calculate generates the following output:



The output is thus:

- Noncentrality parameter  $\lambda$ 
  - In this context, the **noncentrality parameter** is used to measure the power of the  $F$ -test. Larger numbers denote higher possibilities of larger power, depending on the other parameters ( $\alpha$ ,  $\beta$ , &  $N$ ). Although good to report, this isn't critical to consider since it's used to help compute the other values in the Output Parameters section<sup>12</sup>

<sup>12</sup>There may be some use to further explaining the noncentrality parameter statistic: You will see that what is now called the Central and noncentral distributions in the figure do not look normally distributed—especially the red distribution. These distributions are both  $F$ -distributions;  $F$ -tests use  $F$ -distributions to test significance—just like  $\chi^2$  tests use  $\chi^2$  distributions—not normal distributions.  $F$  distributions **strongly resemble**  $\chi^2$  distributions—including looking very non-normal with small  $ns$  (the red distribution in the figure) and becoming more and more normal as the  $n$  increases (the blue distribution in the figure).  $F$ -tests use both the “numerator” degrees of freedom and the “denominator” degrees

- Critical F
  - This is the level of the  $F$ -statistic needed to establish significance under these conditions
- Numerator df
  - This number will be one less than the Number of groups
- Denominator df
  - This is the number that would be needed in the lower part of the  $F$ -test in order to find significance under these conditions—the error degrees of freedom. This in turn translates into most of the sample size we expect to need.
- Total sample size
  - This is the Numerator df plus the Denominator df plus one more degree of freedom needed to estimate the intercept.
- Actual power
  - Given the rounding needed to create a whole number for the numerator and denominator degrees of freedom, actual actual power will often be a little different from what we entered in the Power ( $1 - \beta$  err prob) field in the Input Parameters section.

More about power analyses with one-way ANOVAs is presented by [UCLA's Statistical Methods and Data Analytics site](#).

#### ANOVA: Fixed effects, special, main effects and interactions

(For those who prefer videos—and soothing piano music—[this video](#) also presents conducting sample size estimates for multi-way ANOVAs.)

As we increase the complexity of our analyses, we next move on to ANOVA: Fixed effects, special, main effects and interactions. Here, we can estimate sample sizes for ANOVAs with one or more *nominal* variables.

The main issue with sample size estimates for ANOVA-family analyses is correctly assigning numerator degrees of freedom. And the main issue with doing that when all of the input variables are nominal is to understand how degrees of freedom are computed for them:

- The degrees of freedom for any nominal **main effect** is one less than the number of levels of that variable<sup>13</sup>
  - Using TNM cancer staging as an example, the number of degrees of freedom for its main effect would be  $3 - 1 = 2$ .
- The degrees of freedom for an **interaction** between two nominal input variables is the product of their main effect degrees of freedom.
  - Let us assume that we wanted to look at the interaction between dichotomized gender (self-identifying as female or male) and TNM cancer stage.

---

of freedom, each of these degrees of freedom used to compute its own  $F$  distribution. Since the numerator degrees of freedom are usually quite small (in this example, Numerator df is 2), the  $F$  distribution ushc a small distributions will not look normal. In fact, it can be so non-normal that there is no computable average; the “center” of the distribution is unclear, which is why this is called the are “non-centrality” parameter.

<sup>13</sup>We deduct “1” from each main effect because we only need to establish the values for all levels but one. That last level can be deduced from the other levels. As a simplified analogy, if I knew that  $x + y + z = 6$ , then I only need to know that  $x = 2$  and that  $y = 2$  to know that  $z$  also equals 2.

- \* The degrees of freedom for gender's main effect would be  $2 - 1 = 1$ .
- \* The degrees of freedom for the TNM cancer stage main effect would, of course, be 2.
- \* The degrees of freedom for the gender  $\times$  cancer stage interaction would be  $2 \times 1 = 2$ .

Continuing with that example, if we were interested in knowing the significance of both main effects and their interaction, then the total number of numerator degree of freedom I need to consider in my sample size estimate is:

$$\text{Total Numerator } dfs = df_{\text{Cancer Stage}} + df_{\text{Dichotomized Gender}} + df_{\text{Cancer Stage} \times \text{Gender Interaction}}$$

$$\text{Total Numerator } dfs = 2 + 1 + 2$$

$$\text{Total Numerator } dfs = 5$$

It is thus 5 that we would enter into the Numerator df field.

With this understanding in hand, let us compute the estimate sample size for this model:

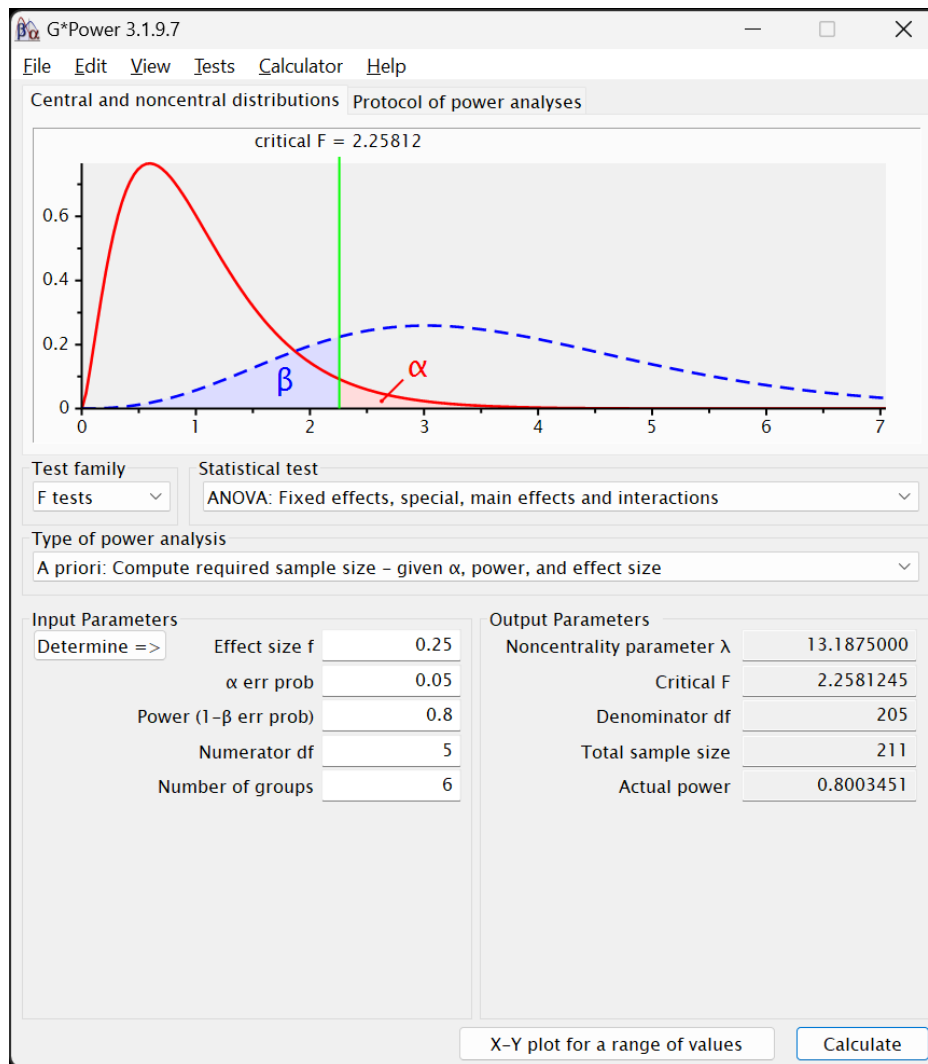
1. Mousing over Effect size  $f$ , we see that a "medium" effect for Cohen's  $f$  is .25, again the default given by G\*Power, and again what we will retain since we have no prior studies to guide us better.
2.  $\alpha$  err prob we will leave as 0.05.
3. Power ( $1 - \beta$  err prob) we will change to .8.
4. Numerator df is 5 (those two main effects and their interaction).
5. Number of groups is the product of the total number of levels of the input variables.

- Here, TNM stage has 3 levels and dichotomized gender has 2, so the Number of groups is  $3 \times 2 = 6$ <sup>14</sup>.

6. Clicking Calculate returns this output:

---

<sup>14</sup>Notice that, in this case, including the interaction term in the model doesn't affect the estimating needed sample size.



Indicating that we expect to need 211 participants to find significant effects for both main effects and their interaction. Of course, we couldn't divide this number evenly between the groups, so there would be a small amount of imbalance between them.

Note that we don't *need* to worry about the significance of every term we enter into a model. As I discuss in Chapter 4, we can enter terms in a model solely to isolate ("partial out") their effects on those other terms whose effects we *are* interested in. Of course, do this cautiously so to not delude yourself into thinking you need fewer participants than you actually will. Sure, data are expensive, but investing to get some but not enough data—and then having to redo a study—is more expensive than getting enough the first time.

## 16.4 R

Among the most—um—powerful packages in R are the `pwr` and `pwr2` packages<sup>15</sup>. `pwr` is suited more for simpler designs (*t*-tests, one-way ANOVAs, etc.) while `pwr2` can handle more complex

<sup>15</sup>[WebPower](#) is also very useful for power analyses.

designs, but both can easily compute same size estimates, etc, with a few succinct lines of code.

### 16.4.1 Comparing Two Independent Correlations

We examine how large a sample is needed to detect a difference between two Pearson correlations:

```
library(pwr)

# Define the correlations
r1 <- 0.5 # The correlation we expect in one group
r2 <- 0.3 # The correlation we expect in the other group

# Compute Cohen's q
q <- abs(0.5 * log((1 + r1)/(1 - r1)) - 0.5 * log((1 + r2)/(1 - r2)))
q

# Estimate required sample size per group to detect difference in correlations
pwr.norm.test(d = q, sig.level = 0.05, power = 0.8,
              alternative = "two.sided")
```

### 16.4.2 Independent Samples t-Test

Sample size needed to detect a medium effect ( $d = 0.5$ ):

```
pwr.t.test(d = 0.5,      # Cohen's d, effect size for a mean difference
           power = 0.8,  # desired power
           sig.level = 0.05, # significance level
           type = "two.sample" # Type of t-test being evaluated
           )
```

### 16.4.3 Paired Samples t-Test

Sample size needed when using paired/matched data:

```
pwr.t.test(d = 0.5,      # Cohen's d, effect size for a mean difference
           power = 0.8,  # desired power
           sig.level = 0.05, # significance level
           type = "paired" # Type of t-test being evaluated
           )
```

### 16.4.4 Point-Biserial Correlation

This is the correlation between a continuous variable and a dichotomous variable (coded as 0 or 1).

```
pwr.r.test(r = 0.3,      # Pearson (or phi) correlation
           power = 0.8,  # desired power
           sig.level = 0.05 # significance level
           )
```

### 16.4.5 One-Way ANOVA

```
pwr.anova.test(k = 3,      # The number of levels of the IV (e.g., 2 for Experimental vs. Control group)
               f = 0.25,   # Cohen's f, the effect size for terms in an ANOVA
               power = 0.8, # desired power
               sig.level = 0.05) # significance level
```

### 16.4.6 Two-Way ANOVA

Two IVs with 3 and 2 levels respectively (6 groups total):

```
pwr2::pwr.2way(a = 3,      # levels in factor A
               b = 2,      # levels in factor B
               alpha = 0.05, # significance level
               power = 0.8,  # desired power
               f = 0.25,    # effect size (Cohen's f)
               n = NULL)    # compute required sample size per cell
```

### 16.4.7 Power Curves

#### 16.4.7.1 Power Curve for Independent-Samples t-Test

```
# Plot the power curve for a range of sample sizes in a two-sample t-test
curve(
  expr = pwr::pwr.t.test(
    n = x,          # sample size per group
    d = 0.5,        # Cohen's d effect size (medium effect)
    sig.level = 0.05, # significance level (alpha)
    type = "two.sample" # specifies independent-samples t-test
  )$power,         # extract power from the result
  from = 10, to = 200, # range of sample sizes (per group)
  xlab = "Sample Size per Group", # x-axis label
  ylab = "Power",             # y-axis label
  main = "Power Curve for d = 0.5" # title of the plot (match d above)
)

# Add a reference line for the conventional 80% power threshold
abline(h = 0.8, col = "red", lty = 2)
```

**16.4.7.2 Power Curve for One-Way ANOVA**

```
## Power Curve for One-Way ANOVA (e.g., 3 groups)

# Plot power vs. total sample size for one-way ANOVA
curve(
  expr = pwr::pwr.anova.test(
    k = 3,                # number of groups
    n = x / k,            # converts total sample size to per-group n
    f = 0.25,             # Cohen's f effect size (medium)
    sig.level = 0.05      # significance level (alpha)
  )$power,
  from = 30, to = 300,    # total sample size range
  xlab = "Total Sample Size", # x-axis label
  ylab = "Power",          # y-axis label
  main = "Power Curve for One-Way ANOVA (k = 3, f = 0.25)" # title
)

# Add conventional power threshold line
abline(h = 0.8, col = "red", lty = 2)
```

**16.4.7.3 Power Curve for Two-Way ANOVA (Main Effects + Interaction)**

`pwr.f2.test()` is used for general linear models including two-way ANOVA. To use this, you need:

- `u`: numerator degrees of freedom (e.g., 1 for each main effect, plus interaction)
- `v`: denominator degrees of freedom (sample size – predictors – 1)
- `f2`: Cohen's  $f^2$  effect size.  $f^2 = \frac{f_{\text{anova}}^2}{1 - f_{\text{anova}}^2}$ , so for  $f = 0.25$ ,  $f^2 \approx 0.0625$

```
## Power Curve for Two-Way ANOVA (2x3 design, 2 main effects + interaction)

# Total df = (levels_A - 1) + (levels_B - 1) + (A*B interaction df)
numerator_df <- (2 - 1) + (3 - 1) + ((2 - 1) * (3 - 1)) # = 1 + 2 + 2 = 5
f2_value <- 0.25^2 / (1 - 0.25^2) # convert Cohen's f to f^2 ≈ 0.0625 / 0.9375 ≈ 0.0667

# Power curve for GLM (e.g., 2x3 ANOVA with 5 df for predictors)
curve(
  expr = pwr::pwr.f2.test(
    u = numerator_df,      # numerator df (main + interaction)
    v = x - numerator_df - 1, # denominator df = N - u - 1
    f2 = f2_value,         # Cohen's f^2 effect size
    sig.level = 0.05       # significance level (alpha)
  )$power,
  from = 60, to = 300,    # total sample size range
  xlab = "Total Sample Size", # x-axis label
  ylab = "Power",          # y-axis label
  main = "Power Curve for Two-Way ANOVA (f = 0.25, df = 5)" # title
)
```

)

```
# Add reference line at power = 0.8
abline(h = 0.8, col = "red", lty = 2)
```

## 16.5 Additional Resources

### 16.5.1 G\*Power Guides & Tutorials

- *G\*Power Manual*, which is quite useful
- The [UCLA Guide to G\\*Power](#) contains detailed but digestible guides to estimates for most of the analyses you'll conduct (except maybe  $\chi^2$  tests).
- Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). [A short tutorial of GPower](#). *Tutorials in Quantitative Methods for Psychology*, 3(2), 51–59. doi: 10.20982/tqmp.03.2.p051.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). [G\\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences](#). *Behavior Research Methods*, 39(2), 175–191. doi: 10.3758/BF03193146. RIS
- Kang, H. (2021). [Sample size determination and power analysis using the GPower software](#). *Journal of Educational Evaluation for Health Professions*, 18, 1–17. doi: 10.3352/jeehp.2021.18.17. RIS

### 16.5.2 Further Readings and Explanations

- Bujang, M. A. (2021). [A step-by-step process on sample size determination for medical research](#). *The Malaysian Journal of Medical Sciences*, 28(2), 15–27. doi: 10.21315/mjms2021.28.2.2
  - Probably best is [Table 1](#) which presents a nice list of other sources for more information about sample size estimations for various types of analyses from correlations to exploratory factor analysis.
- Das, S., Mitra, K., & Mandal, M. (2016) [Sample size calculation: Basic principles](#). *Indian Journal of Anaesthesia*, 60(9), 652–656. doi: 10.4103/0019-5049.190621. PMID: 27729692; PMCID: PMC5037946. NBIB
- Hunt, A. (n.d.). [A researcher's guide to power analysis](#).

### 16.5.3 Sample Size Estimations and Guidelines for More Complex Designs

#### 16.5.3.1 ANCOVAs

- Borm, G. F., Fransen, J., & Lemmens, W. A. J. . (2007). [A simple sample size formula for analysis of covariance in randomized clinical trials](#). *Journal of Clinical Epidemiology*, 60(12), 1234–1238. doi: 10.1016/j.jclinepi.2007.02.006. RIS
- Shieh, G. (2020). [Power Analysis and Sample Size Planning in ANCOVA Designs](#). *Psychometrika*, 85(1), 101–120. doi: 10.1007/s11336-019-09692-3. RIS
- Teerenstra, S., Eldridge, S., Graff, M., de Hoop, E., & Borm, G. F. (2012). [A simple sample size formula for analysis of covariance in cluster randomized trials](#). *Statistics in Medicine*, 31(20), 2169–2178. doi: 10.1002/sim.5352. RIS

### 16.5.3.2 Logistic Regression

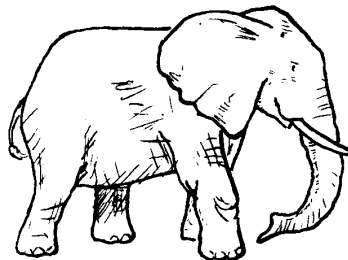
- Motrenko, A., Strijov, V., & Weber, G.-W. (2014). Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics*, 255, 743–752. doi: 10.1016/j.cam.2013.06.031. RIS

### 16.5.3.3 Factor Analysis and Structural Equation Models

- Grace-Martin, K. (n.d.). How big of a sample size do you need for factor analysis? The Analysis Factor. <https://www.theanalysisfactor.com/sample-size-needed-for-factor-analysis/>
- Kelley, K., Lai, K. (2018). Sample size planning for confirmatory factor models. In *The Wiley Handbook of Psychometric Testing* (pp. 113–138). John Wiley & Sons, Ltd. doi: <https://doi.org/10.1002/9781118489772.ch5>. RIS
- La Du, T. J., & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, 74(4), 625–635. doi: 10.1037/0021-9010.74.4.625. RIS
- Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, 5(2), 159–168. doi: 10.1207/s15327574ijt0502\_4. RIS
- Nicolaou, A. I., & Masoner, M. M. (2013). Sample size requirements in structural equation models under standard conditions. *International Journal of Accounting Information Systems*, 14(4), 256–274. doi: 10.1016/j.accinf.2013.11.001. RIS
- Pearson, R., H. & Mundfrom, D. J. (2010). Recommended sample size for conducting exploratory factor analysis on dichotomous data. *Journal of Modern Applied Statistical Methods*, 9(2), 359–368. doi: 10.22237/jmasm/1288584240. RIS
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73(6), 913–934. doi: 10.1177/0013164413495237. RIS

### 16.5.4 Online Tools

- Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., Aczel, B.. (2022). SampleSizePlanner: A tool to estimate and justify sample size for two-group studies. *Advances in Methods and Practices in Psychological Science*. 2022;5(1). doi: 10.1177/25152459211054059. RIS



## Chapter 17

# Introduction to SPSS & Data Preparation

### 17.1 Overview

This document is intended to introduce you to a couple of things. First, we will review SPSS's GUI<sup>1</sup> and how data are prepared and handled therein. We will import a set of rather clean data that we will use to demonstrate ways to further prepare and manipulate data.

Second, we will then introduce some common data exploration functions to understand these data better. We will use this opportunity to further consider some of the concepts we're covering through our other class activities, including normality and outliers.

### 17.2 Orientation to SPSS

#### 17.2.1 Accessing SPSS Through Apporto

SPSS can be accessed online with your CUNY ID through [Apporto](#)—as long as “your browser” is [Chrome](#).

**To access SPSS through Apporto:**

1. Go to CUNY's Apporto login page: <https://cuny.apporto.com/>
2. Enter your CUNY login credentials (your @login.cuny.edu “email” address)
3. If you don't already see an icon for SPSS, in the Apporto [home page](#), click on the App Store button in the top, left corner, just below the hamburger icon that opens up that left-hand menu.
4. Click to Launch SPSS and follow any steps to “optimize<sup>2</sup>” and reconnect.

**To open data in Apporto**, there are two ways:

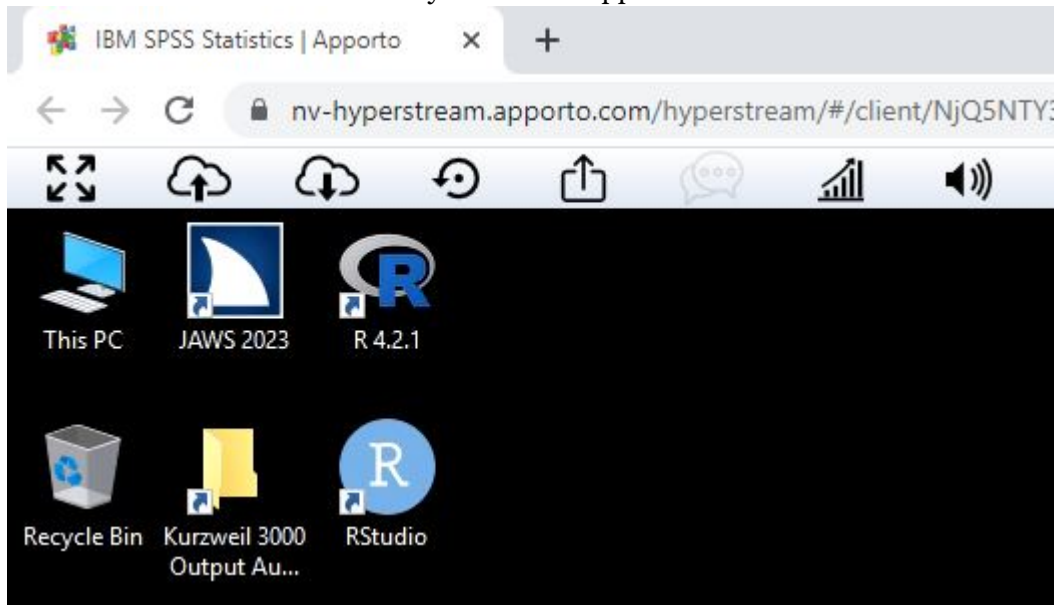
1. Uploading files via dialogue


---

<sup>1</sup>“Graphical user interface”

<sup>2</sup>Because following all of the steps they already laid out for you could not be optimal.

1. Locate the menu bar immediately above the Apporto window:



2. Click on the File upload button (  )
3. Follow the dialogue therein

## 2. Dragging files into the Apporto window

1. Open up a file manager *outside* of the Apporto environment (i.e., in a normal window outside the browser in which Apporto is running)
2. Left click to grab and drag a data file from your file manager into the Apporto window. Apporto will open a notification window letting you know that the file has indeed been imported; it should also now appear in the Apporto window
3. You can then drag the file from Apporto window into the SPSS window that is itself inside Apporto<sup>3</sup>

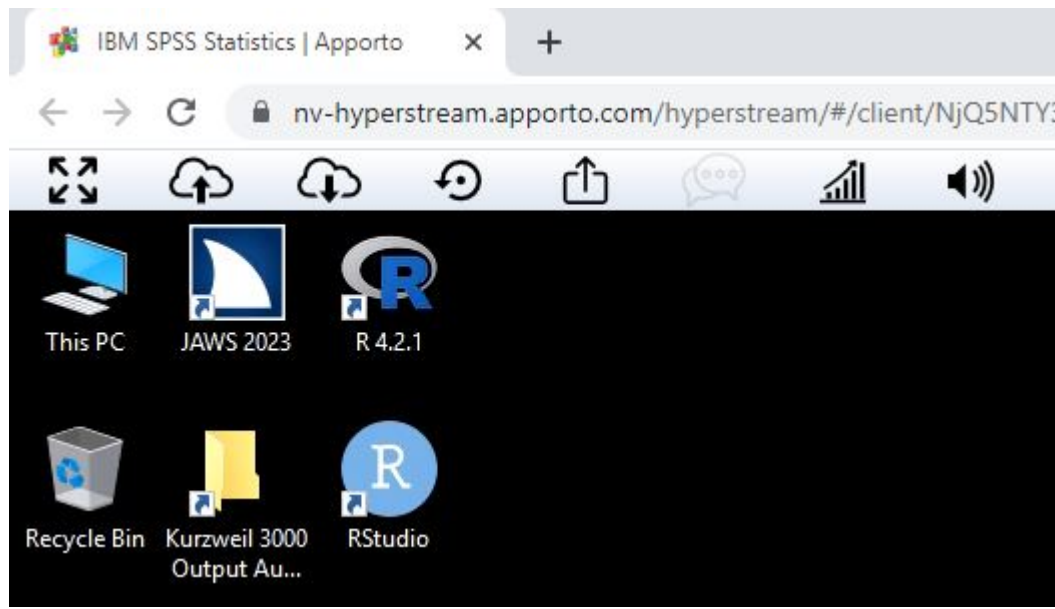
Files you save in Apporto will (at least eventually) appear in **either** the This PC > Desktop folder (accessible from the Desktop folder under Quick Access in Windows' native file manager) **or** in the This PC > Documents folder (Documents under Quick Access).

Clicking on the Settings gear to the right of the Apporto menu bar gives the option to access USBs, although this proved to not always be reliable for all OSs for me.

### To export files from Apporto:

1. In that menu bar immediately above the Apporto window:

<sup>3</sup>You can load it from the Apporto file system via, e.g., This PC > Desktop, but files don't immediately appear there (needing connection refreshes?), so simply dragging it into the SPSS Data Editor window seems most reliable to me.



2. Click on the File download button (  )
3. You get the idea

Alternatively, you can open your email from within Apporto and send it to yourself as an attachment.

### 17.2.2 Editing Global Options

Before we dive into the windows and workings of SPSS, I'd like to note that there are a few useful options to consider modifying given your needs. There are, in fact, many options for tailor SPSS's functioning, output, and performance given throughout its dialogues and within its rather large list of syntax commands. Here, however, we will simply note a few "global" options that can be set to adjust how SPSS acts in general.

To access these, select **Edit > Options** from the menu (in any window). When that dialogue opens, you will see many choices, including the Variable Lists section in the top right of the General tab. In that section, you can choose to either have SPSS default to Display names or to Display labels of variables. As discussed further below, a given variable can be identified by either the shorter, more-restricted name or by the longer label used to describe it. By choosing one of these options you can either show smaller, less-intuitive names or longer, more explanatory labels in (nearly) all of the output SPSS generates. Of course, you can also switch between these as needed.

Some of the other options under the General tab are worth considering (such as whether you want to have SPSS display No scientific notation for small numbers in tables; I mean, we're doing research here, not science). The Language, Viewer, Data, Currency, Charts, Scripts, and Syntax Editor tabs are less useful for most users, but the items in the Output tab's Outline Labeling section may also be worth considering. Either of those options can let you choose whether to show only the variable names, labels, or both; I suggest using Labels for output you share with others, but you may want to use Names for your own analyses since it will make for simpler output.

Under the Pivot Tables tab, you may want to consider changing the TagbleLook to APA\_TimesRoma\_12pt when you're ready to produce pivot tables for your dissertation or publishable manuscripts.

File Locations can be nice to change if you store your data and analyses in dedicated folders.

Finally, you may (or may not) wish to change settings in the Privacy tab.

There is more one can do to customize SPSS output and set defaults that allow for automatic APA styling. Including:

- Using an APA-formatted table to serve as the style for subsequent tables
- Styling a given table using Format >TableLooks
- Styling figures

### 17.2.3 SPSS Windows

SPSS is inherently a syntax-driven program, but its popularity is arguably due in large part to its useful GUI. The GUI has three main windows:

1. The Data Editor which is comprised of the Data View and Variable View tabs
2. The Output window
3. The Syntax Editor window

#### 17.2.3.1 The Data Editor Window

The Data Editor window is the one most commonly used to interface with SPSS. I think one reason for this is that it can help to be looking at one's data while working with it—if nothing else to remember what variables there are and what their names are.

Another reason, though, is because you will have one Data Editor window for each data set you have open; when you access the drop-down menu at the top to, e.g., Analyze your data, SPSS will assume you want to work with the data in whatever window is either currently raised or that was last raised. So, if you have more than one data set open, simply cycle through to the one you want to work with and then choose what you want to be from the drop-down menu—from either the Data Editor, Output, or even Syntax window.

Relatedly, you will notice that the drop-down menu at the top is the same<sup>4</sup> for all of the windows. This indeed means that you don't have to cycle back to the Data Editor window before you do anything. In fact, it can be sometimes easier to use the menu from the Output window so you can look at the results of one command to know what to do with the next. (Anyway, you can see the whole list of variables accessible to a given command in that command's dialogue boxes.)

#### The Data View Tab

The Data View tab<sup>5</sup> presents a spreadsheet of the data. Just like other spreadsheet programs, you can enter, edit, and scroll through your data here. You can use the Page up and Page Down or the arrow keys to scroll. Holding down the Control/Command button while tapping arrow keys will go to the ends of the data; e.g., Control/Command + ↓ will go to the bottom of the data set; Control/Command + ⇒ will go to far right of it, etc. One way this works differently from, e.g., Excel though is that SPSS will skip over empty cells whereas Excel will stop right before each empty cell instead of going all the way to the end.

Right-clicking on things in the Data View tab lets you do some useful things.

<sup>4</sup>Well, actually the Syntax window has a few extra menu items related to running syntax and accessing additional extensions.

<sup>5</sup>The tabs are at the bottom left of the window.

- Right-clicking on a **column header** (i.e., the part at the top that list the variable name) lets you:
  - Sort the entire data set by that variable
  - Copy the variable name or label (more about those things under [Variable View](#))
  - Clear the data set of that variable. **This is the command to delete something in SPSS.** Right-clicking and then choosing Clear will delete the selected cell, row, or column in either the Data View or Variable view tab.
  - Get Variable Information including the variable's name, label, type<sup>6</sup>, any codes for missing values, and the measurement scale for that or any other variable.
  - Send a command to give a nice set of descriptive statistics to the Output window (and go there automatically to see those results)
- Right-clicking on a **row number** lets you:
  - Cut or Copy that row
  - Clear (i.e., delete) that row
  - Insert Cases to manually enter a new row of data (or paste one that you said to cut or copy)
- Right-clicking on a **cell** lets you:
  - Cut or Copy the values in that cell
  - Paste values selected from cutting or copying
    - \* You can also Paste with Variable Names, useful (or confusing) for pasting into a different column
  - Copy the variable name or label
  - Access Variable Information or Descriptive Statistics for that entire variable
  - Clear (i.e., delete) the information in that cell
  - Check the spelling against SPSS's dictionary
  - Change the font slightly




### The Variable View Tab

The Variable View presents what is essentially, a [codebook](#), a list of the variables and information about them, including:

- Name,
  - the name that SPSS uses to access that variable. These are best kept short so that you can see the whole thing in some of SPSS's unnecessarily-small dialogues. They also [can only contain](#) letters, numbers, periods, and underscores.
- Type
  - indicates whether the variable is a String (alphanumeric), Numeric (numbers not specially formatted), or a number with [various types of special formatting](#), such as dates, currency, etc. The Comma and Dot types are for numbers with thousands etc. indicated by commas or dots, respectively<sup>7</sup>. Scientific notation is for numbers formatted like  $1 \times 10^3$  to denote 1,000. Clicking on the button with an ellipsis opens a dialogue where you can change the number type (as well as change the length of the variable—how many characters long it can be).

<sup>6</sup>This "type" is given as either the letter (A or F) or word (DATE, TIME, PCT (for percent), DOLLAR, etc.) followed by a number. An A means that it is a string variable (i.e., Alphanumeric), and an F means it's a number (an "F" is used for [esoteric reasons](#)). The number presents the number of digits possible before and after the decimal point; if the value has no decimal (e.g., F4), then that variable has no decimals.

<sup>7</sup>I.e., Comma is for numbers formatted like 1,000.00 and Dot is for numbers formatted like 1.000,00

- Width,
  - which simply indicates how many characters long or how many digits a variable has left of a decimal. No big deal
- Decimals
  - presents how many decimal places a (numeric) has been assigned.
- Label
  - is very useful. In this field you can write a **rather long** description of what a given variable measures. You can use nearly any characters here to explain it well. To create or change a label, simply left-click inside that field and start typing.
- Values
  - is also quite useful; for variables that are encoded with numbers, you can use this field to indicate what each level of the variable actually denotes. For example, if you have a Likert-style response encoded a number from 1 to 5, you can click on the ellipsis button to denote that 1 = Strongly Disagree, etc. When you explore the variable with descriptives, etc. SPSS will use these value labels instead, making output considerably easier to read. We will show an example of doing this [below](#).
- Missing
  - is yet another useful field. Sometimes a certain character or value will be used to denote a missing value. For example, 99 or NA may be used as place-holders to signify that that datum is actually missing. By clicking on the ellipsis button, you can denote this. We do this [below](#).
- Columns
  - simply notes how many characters wide a column is. You can change the value here or, under the Data View tab, left-click the space between two rows to change this.
- Align
  - just indicates the left, right, or center alignment of a column.
- Measure
  - is an unexpectedly important attribute of a variable. SPSS is quite finicky about the “measure” type of a variable: You can only perform actions on a variable that match that variable type. For example, you can only run correlations on continuous variables. The measurement types that SPSS allows are:
    - Scale denotes a “scalar” variable, which corresponds to either of Steven’s “interval” or “ratio” levels. It is indicated by a little ruler (.
    - Ordinal denotes a, well, ordinal variable and is indicated by a little histogram (.
    - Nominal denotes a nominal variable and is indicated by a cute little Venn diagram (.
- Role
  - is a rather under-utilized field. It can be used to indicate whether a variable is a predictor / independent variable (Input), a outcome / dependent variable (Target), Both, or whether it is used to Partition or Split the data set. We will create a variable that indeed partitions when we [subset the data to only include migrant students](#).

### 17.2.3.2 The Output Window

Another reason I think SPSS is so widely used is because, with just a few mouse clicks, it delivers copious amounts of output. As I noted in class, personally I've found that some researchers use this output to determine their analyses, assuming that if some stat program spits it out, it must be good. Nonetheless, it *can* be good—and certainly makes it worth annotating the output.

#### Annotating Output

The Output window is comprised of two sections, an outline and a main window. The information in either can be changed or added to manually. This can be a good idea. First, of course, because SPSS *does* return a lot of results and sifting through even a few sets of analyses can be tedious.

Second, I strongly recommend taking notes on what you are doing in your analyses and what your thoughts on them are. With data and analyses of any real size and complexity, it can be difficult to jump back in to your analyses even a week or so later; steps that seemed obvious and important at the time can quickly become obscure and lost.

Ways of annotating your output:

- **Insert a heading in the outline** by clicking Insert > New Heading. This will create a new heading at the cursor; double-click on this heading to type in a phrase that will remind you of what you are doing in that section of the output.
  - Alternatively, you can simply double-click on an existing heading to change it. For example, if you conduct more than one *t*-test output, you can double click on the first to change it to *t*-test of *toca.pro* by group and the second to *t*-test of *toca.dis* by group. You can left-click and drag the spacer between the windows to make the outline section wider, but you'll still not want to make the headings too long since they'll quickly become longer than a useful outline window.
- **Insert notes into the output itself** by clicking Insert > New Text. This will create a text box in the output section into which you can write pretty much whatever you want. Unlike a heading, this can be as long as you want to give yourself and your colleagues as much information about what you are doing and what it means.
- You can use the Insert menu to **insert other things**, too, including whole titles for the output, images, etc.

Note that you can also double-click on any element in the main output section to manipulate that element. This way, you can modify the colors, fonts, or even the text within tables, figures, etc.

Of course, you can then save your output (to a .spv file) as notes on your analyses.

#### Exporting Output

Right-clicking on an element lets you copy it to then paste it into, e.g., your manuscript (as we will do in Chapter 3: Writing Results).

Alternatively, you can Export an element. When you right-click on an element and choose to do that, you will be able to export it as a .html, .pdf, .ppt, .doc, etc. For importing into, e.g., Word, I suggest exporting as an .html file.

## Syntax in Output

SPSS is a powerful stats program, but I personally think that its GUI is a big reason for its success. Nonetheless, SPSS's GUI is in fact just an “overlay” that just lets us access its most common commands more intuitively; SPSS is in fact running the syntax that those mouse clicks created.

SPSS versions 27 and earlier return the syntax it used to generate results in the Output window by default right above the given results<sup>8</sup>. As of version 28, it does not. We can set SPSS to automatically return the syntax used in the output by going to Edit > Options > Viewer and then checking the Display commands in the log box in the lower-left of that Viewer window<sup>9</sup>.

Why do this? Because there are several ways in which the syntax that SPSS posts can be quite useful. First, you can copy that syntax into the Syntax Editor (as noted **below**) to rerun any analyses. This is useful when you are returning to analyses later on and, e.g., want to generate a smaller set of analyses.

Second, as you learn what SPSS can do, you can use the syntax to learn better *how* to do it—and how to tweak your analyses to get exactly the output you want. Reviewing existing syntax is a lot easier than learning it from scratch.

Third, once you've gained some facility using SPSS, you will find that there are things you want to do that you can't through the GUI. Instead, you will need to do things directly with the syntax. Although you certainly can type syntax directly into the Syntax Editor, it's often easier to paste in existing syntax and edit it as needed. In fact, in the long run, that's also faster.

Fourth, you can annotate syntax a bit like you can annotate output. This way, you can create and save a syntax file (saved as a .sps file) that's a *lot* smaller and easier to navigate through than some massive output file—and still be able to generate that mountain of results with a few quick keystrokes<sup>10</sup>.

### 17.2.3.3 The Syntax Window

SPSS doesn't open a Syntax window automatically, like it does a Data Editor or Output window, but simply clicking File > New > Syntax opens one. We will demonstrate using it **below**, but the **general way** to use it is to either paste in or type some syntax command and, with the cursor in some part of that syntax, either click on the big, green play button<sup>11</sup> or type Control/Command + R.


SPSS syntax itself follows a set grammar. Some command is given first; often this is immediately followed by a “statement” that just tells SPSS what variables, etc. to run that command on. This is followed by one or more options, for example whether to print out both figures and tables based on the command. Critically, each command must end with a period.

As you might expect, SPSS has **many** commands to choose from; **more** are available if you pay them more (and have your own copy of SPSS; this won't work with the version we have access to through CUNY).

<sup>8</sup>The syntax is posted under Log headings in the outline. This is useful for finding it, but the log is also used by SPSS to report errors and warnings, so it can be a little confusing to find the syntax or even know that errors/warnings were generated.

<sup>9</sup>We can also turn on outputting syntax with *syntax*: SET PRINTBACK LISTING. turns it on, and SET PRINTBACK NONE. turns it off.

<sup>10</sup>Control/Command + A to select all of the syntax in the window, and then Control/Command + R to run it all.

<sup>11</sup>I.e., this button: 

## 17.3 Data Preparation & Cleaning

This section will use selections for the publicly-available data from the University of North Carolina at Chapel Hill's National Longitudinal Study of Adolescent to Adult Health ([Add Health](#)) study (stored on the University of Michigan's [ICPSR](#) repository). This study “is a longitudinal study of a nationally representative sample of over 20,000 adolescents who were in grades 7-12 during the 1994-95 school year, and have been followed for five waves to date, most recently in 2016-18. Over the years, Add Health has collected rich demographic, social, familial, behavioral, psycho social, cognitive, and health survey data from participants and their parents ... [including] data from participants' schools, neighborhoods ... and in-home physical and biological data.”

Please access that the selection of data we will use here from:

- **Add Health Data Mostly Ready**

After downloading that set of data, please upload them into SPSS (e.g., via Apporto, Section 17.2.1).

These data are indeed nearly ready for further analyses, but have a few issues to address to demonstrate how to clean or improve data in ways that are commonly needed.

Note that we will be using these data for some future activities, so you may want to save them after you've taken the steps below to fully clean them.

### 17.3.1 Changing the Measurement Level of a Variable

The AID variable is right now a Scale variable; it is a number after all. And it's not uncommon for SPSS to import IDs as numbers since replacing names with numbers is a very typical way to anonymize participants. And leaving it as a number (a Scale level Measure) won't likely create any problems in SPSS<sup>12</sup>, but it still presents a good opportunity to demonstrate changing the Measure of a variable. To do this:

1. Go to the Variable View tab of the Data Editor window.
2. Left-click on the Measure cell in the AID variable's row. When you do, a drop-down menu will appear listing the three measure levels.
3. Select to make AID a Nominal variable.

Now, SPSS will “understand” that this is in fact a name that signifies each participants and should be treated as such in all analyses.

### 17.3.2 Creating a Variable Label for a Variable

Continuing to prepare AID, let's now give it a variable label. SPSS requires that variable names (in the Names column) be relatively brief<sup>13</sup> and only use certain characters<sup>14</sup>. In fact, it's often

<sup>12</sup>As we'll discuss briefly in the measurement class, interval and ratio variables—those that SPSS calls Scale variables—can be analyzed in more ways than ordinal variables; ordinal, in turn, can be analyzed in more ways than nominal.

<sup>13</sup>SPSS variable names can be up to 64 characters long.

<sup>14</sup>SPSS variable names can include letters, numbers, periods, and underscores (\_). They must also begin with a letter.

good to keep them short since too-long variable names can be hard to read in the tiny windows SPSS uses for most dialogues<sup>15</sup>.

Variable labels (in the Label column) can be much longer<sup>16</sup> and contain many more types of characters<sup>17</sup>. They do not work well in the SPSS dialogues, but are often great for tables and figures.

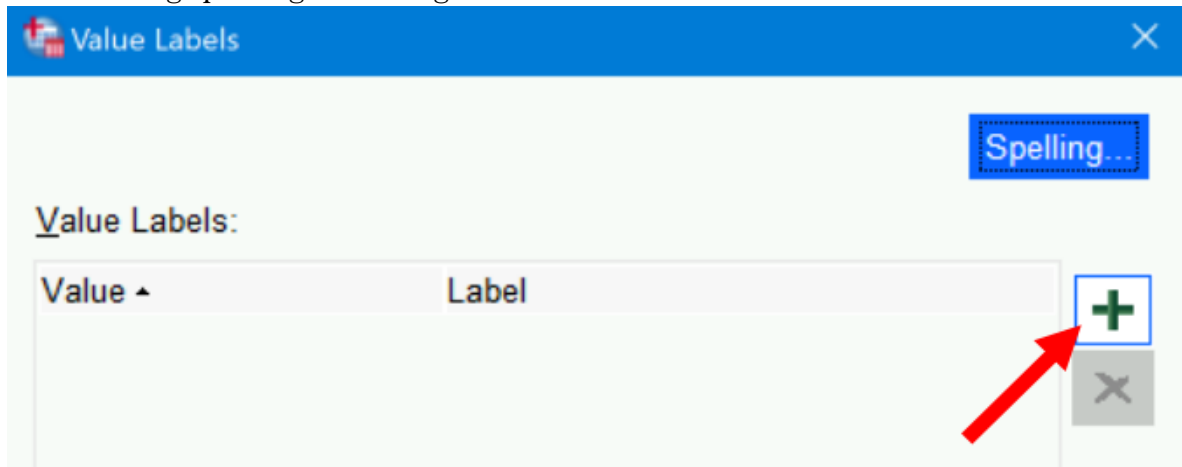
To add a label to AID simply:

1. Single-left click in the Label cell for AID, and type/paste: Unique Participant ID or something like that.

### 17.3.3 Setting Values Labels for a Variable

In addition to giving a more human-friendly label to a variable, we can give clearer labels to the levels of a variable. These can also be shown in figures and tables, making those easier to understand and helping avoid misinterpretation.

1. Also in the Variable View of the Data Editor, click on the Values cell in the the Bio\_Sex row.
2. Click on the ellipsis button that appears.
3. In the dialogue box that opens, enter a 0 in the Values field.
4. Click the large plus sign to the right of the Value Labels field:

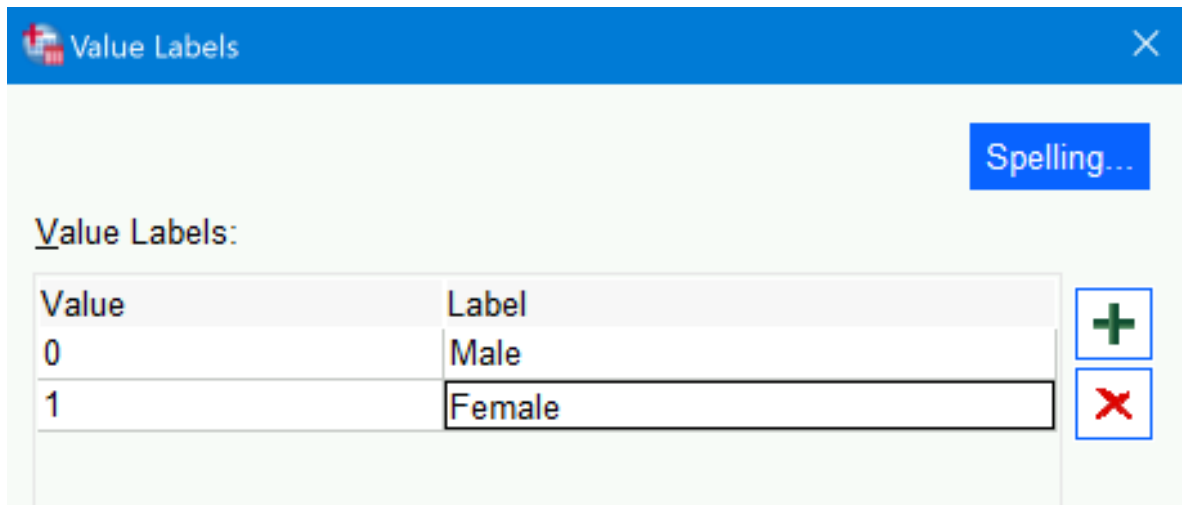


5. Type Male in the Label field.
6. Click the plus sign again. 0 and Male now appear in the field next to the Add button, and an other row below that has appeared.
7. In that second row, type a 1 in the Values field and Female in the Label field:

<sup>15</sup>We can change whether we see variable names or labels in dialogues via Edit > Options; under the General tab, go to the Variable Lists section near the top left; there, select either Display labels or Display names.

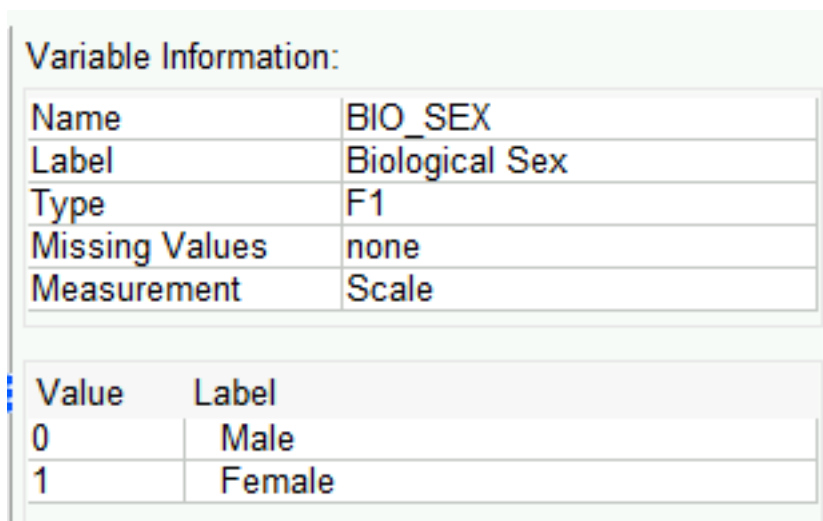
<sup>16</sup>SPSS variable labels can be up to 256 characters long.

<sup>17</sup>SPSS variable labels can contain nearly any printable character including spaces, punctuation, and even emojis. !



8. Again click the Add button to add this association as well.
9. Click OK

Now when you click on the values cell for the Bio\_Sex row, you will see these value labels added. Right-clicking on the Bio\_Sex row and choosing to look at the Variable Information will show these in addition to the other information:



Note that we have not actually changed the data. They are still numbers (Scale level measures). Right-click again on that variable (in either the Data View or Variable View tabs) and select Descriptive Statistics. You will see in the output that SPSS generates means, etc. just as it would for any interval/ratio variable:

## Statistics

### Biological Sex

N	Valid	18289
	Missing	2
Mean		.53
Median		1.00
Std. Deviation		.499
Range		1
Minimum		0
Maximum		1

However, now in the drop-down menu click on Analyze > Descriptive Statistics > Frequencies and you will see that the level values are replaced with the more explanatory value labels, helping us (and out colleagues and readers) more easily see what the responses really meant:

### 17.3.4 Setting Missing Values

As mentioned briefly above, we can set certain values to be recognized as representing missing values. Most of the variables in this set were imported with blank cells denoting missing values or missing values already established. However, running descriptives<sup>18</sup> on Weight shows that the maximum weight (in kgs) here is 9999. Since your momma was not a participant, that value is surely intended to denote missing values.

We can easily fix this:

1. In the Variable View tab of the Data Editor window, click on the the ellipsis button in the Missing cell of the Weight row
2. Click on the radio button next to Discrete missing values
3. In the first field under that, type in 9999
4. Click OK

### 17.3.5 Create Dummy Variables from Different Levels of a Nominal Variable

I am an advocate for using [dummy variables](#). They can make it easier to interpret the effects of each level of a nominal variable without needing to resort to, e.g., post hoc analyses.

Dwelling\_Type is coded right now as a numeric variable. Clicking on the ellipsis button in the Values column for that variable presents the following definitions for the values:

<sup>18</sup>Again, you can either right-click on that variable in the Data Editor and select Descriptive Statistics or go to Analyze > Descriptive Statistics > Descriptives in the drop-down menu.

Value	Label
1	(1) Detached single-family house
2	(2) Mobile Home/trailer
3	(3) Single-family row/town house (2 or more attached units)
4	(4) Divided house
5	(5) Small apartment building (2-4 units)
6	(6) Apt building (5 or more units)/free access to housing un[it]
7	(7) Apt building (5 or more units)/locked entry/doorman/both
8	(8) Other

We could leave this as single variable; for ANOVAs this may help since then we would only look for differences between these levels if the ANOVA found a significant main effect for this variable<sup>19</sup>. However, it is more flexible and efficient for other types of models to convert these levels into meaningful dummy variables that can be added as needed. (By not necessarily including all levels—all dummies—we could also save a few degrees of freedom, too.)

### 17.3.5.1 Quickly Creating Dummy Variables for Variable Levels

We could simply convert each level into a separate dummy variable. To do this:

1. Under the Transform menu, select Create Dummy Variables
2. Select Dwelling\_Type under Variables and then add that to the Create Cummy Variables for: field by again clicking on the arrow (→)
3. We are going to create a simple dummy variable—not, e.g., one derived from a combination of other variables—so leave Create main-effect dummies selected
4. It's fine to leave selected Use value labels under Dummy Variable Labels since neither choice matters for a simple “main effect” dummies
5. Under Macros, select to Omit first dummy category from macro definitions. We can nearly always select to do this because we usually need one fewer dummy variables than there are values in the original variable. The Population variables has two values (Migrant and Non-Migrant), so we only need one dummy variable (i.e.,  $2 - 1 = 1$ ) to fully encode the information in the Population variable<sup>20</sup>  
This will make Detached single-family the “reference” group: If we included all of the dummy variables we're creating in a model, then their effects would be relative to those living in single-family detached homes.
6. In the Root Names field, type Dwelling. This will add that word to the each of the dummy variables to remind us where they came from and what they're referring to.
7. Click OK

Dummy variables can only take on the values of 0 or 1. For some reason, SPSS gives dummies it creates two decimal places. We clearly don't need these, so:

1. In the Variable View tab, click into the Decimals cell of the population\_1 variable<sup>21</sup>
2. Change the value to 0

Note that we could also change to Width to 1 since we only need one digit to the left of the decimal.

<sup>19</sup>Remember that ANOVAs conduct an “omnibus” *F*-test to first find if there is any significant difference anywhere between the levels. We then conduct post hoc analyses to investigate where those differences are.

<sup>20</sup>Note that SPSS may create two variables anyway. I'm not sure why it does this, but we can simply delete (Clear) the one with the Population=Non-Migrant label since we'll only work with the migrant students.

<sup>21</sup>Or whichever is the dummy with the Population=Migrant label that we'll be keeping.

### 17.3.5.2 Creating Dummy Variables for Combined Variable Levels

Some of these dwelling types are quite similar to each other; for at least preliminary analyses, then, we will combine some of them into the same dummy variable<sup>22</sup>

We could group these several ways, of course, but let's group them thusly:

- Detached\_House: 1 if the variable value is 1, else it will be 0
- Mobile\_or\_RowHouse: 1 if the variable value is either 2 or 3, else 0
- MultiUnit\_Housing: 1 if the value is 4 – 7, else 0
- Other\_Dwelling: 1 if the value = 8, else 0

We will do this by using an other type of data transformation. This process is not as straightforward as a batch creation of the dummies, but still not onerous. It's also useful for many other types of transformations—not just into dummy variables:

1. Under the Transform menu, select Compute Variable
2. In the Target Variable box, type Detached\_House
3. In the Numeric Expression box, type: (Dwelling\_Types = 1)
4. Click OK to create the variable. You now have a new variable Detached\_House coded as:
  - 1 if the respondent lives in a detached single-family house
  - 0 for all other types of dwellings
5. Repeat Steps 1 – 4 to create the second dummy variable, using:
  - For Target Variable type Mobile\_or\_RowHouse
  - For Numeric Expression type (Dwelling\_Types = 2 OR Dwelling\_Types = 3)  
This variable will equal:
    - 1 if the respondent lives in a **mobile home/trailer** or a **row/town house**
    - 0 otherwise
6. Now repeat Steps 1 – 4 again to create the third dummy variable, using:
  - For Target Variable type MultiUnit\_Housing
  - For Numeric Expression type (Dwelling\_Types >= 4 AND Dwelling\_Types <= 7)  
This captures all multi-unit dwellings:
    - Divided house
    - Small apartment buildings
    - Larger apartments with or without doormen / locked entries
7. One more time, repeat to create the final dummy:
  - For Target Variable type Other\_Dwelling
  - For Numeric Expression type (Dwelling\_Types = 8)  
This dummy equals 1 only if the dwelling type is coded as Other

After creating the variables, go to Variable View to create variable labels and perhaps labels for the levels.

---

<sup>22</sup>Of course, if it later turned out that this combined dummy variable was important to unpack we easily could by then creating separate dummies for them.

### 17.3.6 Transform an Old Variable into a New One

It's often necessary or useful to create a new variable based on the values of one or more existing variables. For example, we may want to recode a variable or to compute an overall score.

A common reason to recode variables is to “reverse” score them—switch the direction of the values so they match either what is more intuitive or to work better with other variables in your data. In other words, to change the values like this:

Table 17.2: Example of Reverse Scoring Items

Participant's Response	Original Values	New Values
Most/all of the time	0	3
A lot of the time	1	2
Sometimes	2	1
Never/rarely	3	0

Among the variables selected from the Add Health data are a series that ask about the participants feelings during the past week. The prompts follow this template:

“How often was each of the following things true during the past seven days?: You were \_\_\_\_.”

Most of these items ask about negative emotions/affects, things like feeling sad or lonely. A few, however, ask about positive emotions/affects.

Right now, the position-affect items are scored like the Original Values in Table 17.2, just above. This allowed us to create the Overall\_Negative\_Affect score, which is simply the average of all of a participant's responses to those items; higher numbers on this variable denote more *negative* emotions/affects. Since the positive affect items are part of this score, it was easiest to have them coded in the opposite direction from the negative affect items; the Overall\_Negative\_Affect score thus represents both more negative affects and fewer positive ones.

It may also be useful to create an Overall\_Positive\_Affect score—and to code it so that higher values denote more positive affect. One way to do this would be to first reverse score those positive affect items and then to compute an overall score for them.

#### 17.3.6.1 Recode Variables

We will reverse score the positive affect items near the end of the Add Health dataset<sup>23</sup>.

1. Click on Transform > Recode into Different Variables...<sup>24</sup>

<sup>23</sup>We could also reverse code by subtracting the variable from a constant. Here, we could subtract from 3. This would make an initial 2 into  $3 - 2 = 1$ , etc.

Often, though, we want the numbers to range from 1 to the maximum. In that case, we subtract the variable from 1 greater than the max value. If, e.g., instead we had variables that ranged from 1 to 4, we could reverse code them by subtracting from 5:

$5 - 1 = 4$

$5 - 2 = 3$

$5 - 3 = 2$

$5 - 4 = 1$ .

<sup>24</sup>You may Recode into Same Variables...—and sometimes I do—but this can be dangerous if you make a mistake or tedious if you decide you want the original scaling back. Recoding into a different variable also makes it clearer that you indeed did just that in case you come back later and don't quite remember if you did or not.

2. In the dialogue that opens, move Happy, Enjoyed\_Life, Just\_as\_Good, and Hopeful to the Numeric Variable -> Output Variable field.
  3. Left-click on the first row in the Numeric Variable -> Output Variable field, thus selecting Happy --> ?.
  4. In the Output Variable section just to the right, type Happy\_Reversed in the Name field and—if you want—How Often Felt HAPPY Past Week - Reverse Scored in the Label field. Click the Change button under those fields in the Output Variable section.
- The first row in the Numeric Variable -> Output Variable field now shows that Happy is going to be recoded into a new Happy\_Reversed variable:

5. Now select the Enjoyed\_Life --> ? and type Enjoyed\_Life\_Reversed in the Name field, How Often ENJOYED LIFE Past Week - Reverse Scored in the Label field before again clicking the Change button.
- Continue with creating a Just\_as\_Good\_Reversed (labeled How Often Felt JUST AS GOOD AS OTHER PEOPLE Past Week - Reverse Scored) and Hopeful\_Reversed (labeled How Often Felt HOPEFUL ABOUT THE FUTURE Past Week - Reverse Scored) variables.
6. Now, click on the Old and New Values button underneath the Numeric Variable -> Output Variable field section.
  7. In the dialogue that opens, type 0 in the Value field in the Old Value section and 3 in the Value field in the New Variable section. Next, click the Add button in the Old -> New: section under the the New Variable section. 0 --> 3 will now appear in there:


8. Continue setting up the recoding by adding 1 --> 2, 2 --> 1, and 3 --> 3 into the Old --> New: section.
9. To ensure any other values are addressed (e.g., 999 for missing values), click the All other values radio button at the bottom of the Old Value section and the System missing button in the New Value section; Add that as well to the Old --> New: section:

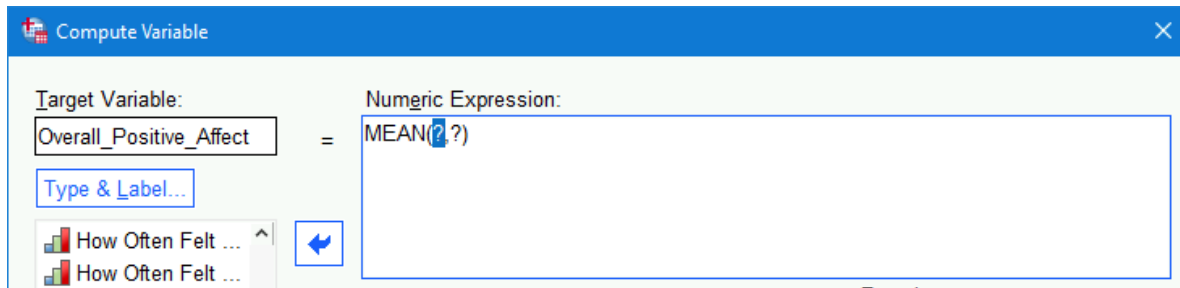
10. Click Continue and then OK in the original dialogue box.

Note that the variables are created as nominal. This won't affect the score computation in 17.3.6.2, but will need to be changed for proper analyses otherwise.

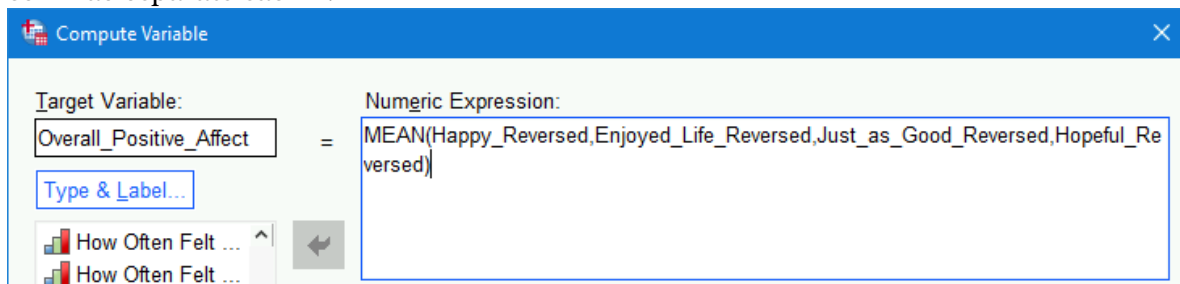
### 17.3.6.2 Compute an Overall Score From Several Variables

We will compute an Overall\_Positive\_Affect score from those four variables we just created. We will compute it from the average of those four variables, however we could instead sum them. Taking the average, though, keeps the score on the same scale as the variables that comprise it, which can make interpreting and comparing them easier. (Of course, standardizing them all does that even better.)

1. Click on Transform > Compute Variable.
2. Type Overall\_Positive\_Affect in the Target Variable field in the top left of the dialogue that opens.
3. Under the Functions and Special Variables section in the bottom right of that dialogue, click on Mean.
4. Next, click on the “up” arrow () next to that section.
5. MEAN(?,?) will now appear in the Numeric Expression: field at the top:



6. Select Happy\_Reversed from the list of variables and click on the right arrow to add that where the first ? appears in the Mean(?,?) formula.
7. Next, add Enjoyed\_Life\_Reversed to the formula.  
Note that the values within the parentheses of the Mean(?,?) formula need to be separated by commas, so make sure there is a comma separating it from Happy\_Reversed.
8. Continue to add Just\_as\_Good\_Reversed and Hopeful\_Reversed to the formula, ensuring that commas separate each<sup>25</sup>:



9. Click OK to finish.

### 17.3.7 Standardize Variables

I'm a big fan of standardized data. It doesn't change the distribution of scores at all but makes values on one variable directly comparable to values on an other—even if they're measured on very different scales<sup>26</sup>

SPSS makes it very easy to standardize variables:

1. Click on Anlyaze > Descriptive Statistics > Descriptives
2. Click on Waist\_Circum
3. Now, *holding down the Shift key* either single-(left-)click on EBV or tap the down-arrow key until you have selected all of the variables from Waist\_Circum to EBV
4. Now click on the blue arrow to move all of those variables to the variabls field
5. Under the Options dialogue, we might as well check to review, e.g., the Mean, Std. Deviation, Minimum, Maximum, and S. E. Mean (i.e., the **standard error of the mean** that we covered in our **first lecture**)
6. But our real goal here is to check the Save standardized values as variables before clicking on OK

And that's all it takes to create standardized variables. They will now all appear at the end of the dataset. The variable names will be like the original, pre-pending with a Z, e.g., Waist\_Circum becomes ZWaist\_Circum.

<sup>25</sup>You can also type into that Numeric Expression: field, so here you can simply paste this into that field:  
MEAN(Happy\_Reversed,Enjoyed\_Life\_Reversed,Just\_as\_Good\_Reversed,Hopeful\_Reversed)

<sup>26</sup>In analytic models—if all variables are either standardized or dummy-coded—it also lets us remove the intercept term, making our analyses a bit more powerful.

## 17.4 Export a Table in APA Format to Word

To export a table in APA format in SPSS, you can use the following steps. This process involves generating the table, modifying it to meet APA style guidelines, and then exporting it.

### 17.4.1 Steps to Export a Table in APA Format in SPSS:

#### 1. Generate the Table:

- First, create the table you need by running the appropriate analysis.
- For example, to create a descriptive statistics table, go to Analyze > Descriptive Statistics > Descriptives..., select the variables, and run the analysis.

#### 2. Modify the Table:

- Once the table is generated, it will appear in the Output Viewer.
- To modify the table, double-click on it to open it in the Pivot Table Editor.
- Adjust the table's appearance to match APA style as closely as possible. This might include:
  - Ensuring that the table uses a simple grid with minimal lines.
  - Aligning text correctly (typically left-aligned for text, right-aligned for numbers).
  - Using appropriate font and size (Times New Roman, 12-point is common for APA).
  - Including relevant statistics (e.g., means, standard deviations).

#### 3. Export the Table:

- After making the necessary adjustments, you can export the table.
- Click on File > Export... in the Output Viewer.
- In the Export Output dialog box, choose the desired file format. For APA tables, Microsoft Word (.doc or .docx) is typically a good choice.
- Specify the file name and location where you want to save the file.
- Under Objects to Export, select All Visible Objects or choose the specific table you modified.
- Click OK to export the table.

### 17.4.2 Example Export in Word:

#### 1. Open the Exported File:

- Open the exported Word document.
- Review the table to ensure it adheres to APA formatting guidelines.

#### 2. Adjust in Word if Necessary:

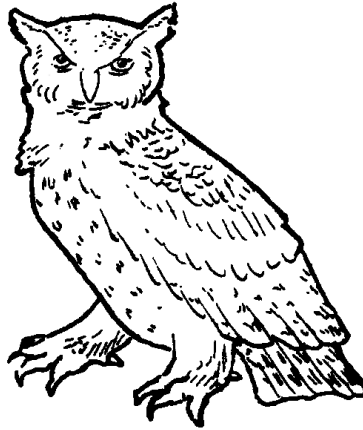
- If further adjustments are needed, you can make them directly in Word.
- Ensure the table is labeled correctly with a table number and title (e.g., Table 1).
- Include any notes below the table, formatted according to APA guidelines.

### 17.4.3 Summary

To export a table in APA format from SPSS, generate the table through the appropriate analysis, modify it using the Pivot Table Editor to adhere to APA style, and then export it to a Word document. Make any final adjustments in Word to ensure the table fully complies with APA formatting standards.

## 17.5 Additional Resources

- Supplemental materials from Polit & Beck (2017)
  - *SPSS Analysis of Descriptive Statistics*
  - *SPSS Analysis of Inferential Statistics*
  - *SPSS Analysis and Multivariate Statistics*



## Chapter 18

# Data Exploration with R

This chapter is among the “stubs” I’m slowly working on. There isn’t much in this one, but still enough that I figured it was worth making it public.

### 18.1 Common Exploration Commands

<https://www.r-bloggers.com/2018/11/explore-your-dataset-in-r/>

Nearly the opposite of SPSS, R is a “quiet” language that only talks back to you when you explicitly ask it to. This isn’t always desirable when you’re exploring data since there may be aspects of the exploration you would have forgotten or not thought of doing if you hadn’t seen the output first. At least R makes up a bit for it with some pretty pithy commands—and even a few lengthy outputs—that we’ll cover here.

We’ll use some pre-existing data packaged with R for our examples here. Which data are available can be easily found with:

```
data(infert)
head(infert)
```

```
education age parity induced case spontaneous stratum pooled.stratum
1 0-5yrs 26 6 1 1 2 1 3
2 0-5yrs 42 1 1 1 0 2 1
3 0-5yrs 39 6 2 1 0 3 4
4 0-5yrs 34 4 2 1 0 4 2
5 6-11yrs 35 3 1 1 1 5 32
6 6-11yrs 36 4 2 1 1 6 36
```

```
summary(infert)
```

```
education    age      parity    induced
0-5yrs : 12 Min. :21.00 Min. :1.000 Min. :0.0000
6-11yrs: 120 1st Qu.:28.00 1st Qu.:1.000 1st Qu.:0.0000
```

```

12+ yrs:116 Median :31.00 Median :2.000 Median :0.0000
      Mean :31.50 Mean :2.093 Mean :0.5726
      3rd Qu.:35.25 3rd Qu.:3.000 3rd Qu.:1.0000
      Max. :44.00 Max. :6.000 Max. :2.0000
case      spontaneous      stratum      pooled.stratum
Min. :0.0000 Min. :0.0000 Min. :1.00 Min. :1.00
1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:21.00 1st Qu.:19.00
Median :0.0000 Median :0.0000 Median :42.00 Median :36.00
Mean :0.3347 Mean :0.5766 Mean :41.87 Mean :33.58
3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:62.25 3rd Qu.:48.25
Max. :1.0000 Max. :2.0000 Max. :83.00 Max. :63.00

```

```
dplyr::glimpse(infert)
```

Rows: 248

Columns: 8

```

$ education <fct> 0-5yrs, 0-5yrs, 0-5yrs, 0-5yrs, 6-11yrs, 6-11yrs, 6-11y~
$ age       <dbl> 26, 42, 39, 34, 35, 36, 23, 32, 21, 28, 29, 37, 31, 29, ~
$ parity    <dbl> 6, 1, 6, 4, 3, 4, 1, 2, 1, 2, 2, 4, 1, 3, 2, 2, 5, 1, 3~
$ induced   <dbl> 1, 1, 2, 2, 1, 2, 0, 0, 0, 0, 1, 2, 1, 2, 1, 2, 2, 0, 2~
$ case      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ spontaneous <dbl> 2, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1~
$ stratum   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
$ pooled.stratum <dbl> 3, 1, 4, 2, 32, 36, 6, 22, 5, 19, 20, 37, 9, 29, 21, 18~

```

```

# skimr::skim(infert)
# DataExplorer::create_report(infert)

```

The `data()` command can be used both to load data into R and—with the parentheses left blank—list out whatever data are currently available. Please note that `data()` will list data sets that you loaded in addition to the ones that came pre-installed with R (or any packages you’ve invoked).

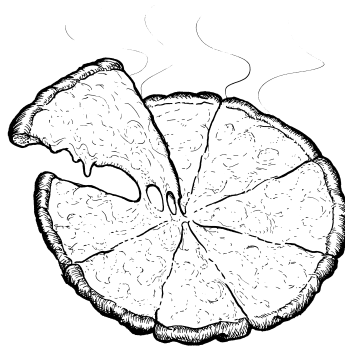
This [Statistics Globe](#) page provides several R commands that are handy for data exploration. More methods for visualizing the data while you explore it are given [this Towards Data Science](#) page.

## 18.2 Using SQL and tidyverse

This [R Views](#) newsletter post by [Vachharajani](#) presents a nice overview using [SQL](#) and the [tidyverse](#) “ecosystem” of packages for R.

[SQL](#) is a venerable programming language used to manage and manipulate data—especially very large sets of data. R uses RAM to hold and manipulate data, and so can flounder with very large sets of data; using SQL can thus help. There are [several ways](#) to use SQL and R together, however the most common are either to first prepare the data in SQL before [exporting it](#) (or parts of it) into R or working from within R to [make queries to the SQL-prepared data from within R](#).

[tidyverse](#) is a set of packages designed to make common tasks—especially the [manipulation](#) and [presentation](#) of data—both more flexible and intuitive. Intuition is a relative thing, and as much as the tidyverse grammar and even vocabulary do make sense, they take some [learning to understand](#)—learning that is really in addition to learning core R syntax and grammar.





# References

- Abelson, R. P. (1995). *Statistics as Principled Argument*. Lawrence Erlbaum Associates Publishers. [https://articles.viriy.net/statistics\\_as\\_principled\\_argument.pdf](https://articles.viriy.net/statistics_as_principled_argument.pdf)
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to Meta-Analysis* (1. Aufl., p. xxix). Wiley. <https://doi.org/10.1002/9780470743386>
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. [https://articles.viriy.net/statistical\\_modeling\\_the\\_two\\_cultures.pdf](https://articles.viriy.net/statistical_modeling_the_two_cultures.pdf)
- Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, 39(4), 860–864. <https://doi.org/10.1080/03610911003650383>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Davis, L. L., Broome, M. E., & Cox, R. P. (2002). Maximizing retention in community-based clinical trials. *Journal of Nursing Scholarship*, 34(1), 47–53. <https://doi.org/10.1111/j.1547-5069.2002.00047.x>
- Gustavson, K., von Soest, T., Karevold, E., & Røysamb, E. (2012). Attrition and generalizability in longitudinal studies: Findings from a 15-year population-based study and a Monte Carlo simulation study. *BMC Public Health*, 12, 918. <https://doi.org/10.1186/1471-2458-12-918>
- Hausman, C., & Rapson, D. S. (2017). Regression discontinuity in time: Considerations for empirical applications. *National Bureau of Economic Research Working Paper Series*, No. 23602. <https://doi.org/10.3386/w23602>
- Hausman, C., & Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10(1), 533–552. <https://doi.org/10.1146/annurev-resource-121517-033306>
- Kao, L. S., & Green, C. E. (2008). Analysis of variance: Is there a difference in means and what does it mean? *Journal of Surgical Research*, 144(1), 158–170. <https://doi.org/10.1016/j.jss.2007.02.053>
- Khamis, H. (2008). Measures of association: How to choose? *Journal of Diagnostic Medical Sonography*, 24(3), 155–162. <https://doi.org/10.1177/8756479308317006>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Monsalves, M. J., Bangdiwala, A. S., Thabane, A., & Bangdiwala, S. I. (2020). LEVEL (Logical Explanations & Visualizations of Estimates in Linear mixed models): Recommendations for reporting multilevel data and analyses. *BMC Medical Research Methodology*, 20, 1–9. <https://doi.org/10.1186/s12874-019-0876-8>
- Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, 40(2), 129–147. <https://doi.org/10.2333/bhmk.40.129>
- Polit, D. F., & Beck, C. E., Cheryl T. (2017). *Nursing Research: Generating and Assessing Evidence for Nursing Practice* (Tenth). Wolters Kluwer.
- Raper, S. (2020). Leo Breiman's "two cultures". *Significance*, 17, 34–37. <https://doi.org/10.1111/j.1740-9713.2020.01357.x>

- Teague, S., Youssef, G. J., Macdonald, J. A., Sciberras, E., Shatte, A., Fuller-Tyszkiewicz, M., Greenwood, C., McIntosh, J., Olsson, C. A., & Hutchinson, D. (2018). Retention strategies in longitudinal cohort studies: A systematic review and meta-analysis. *BMC Medical Research Methodology*, 18(1), 151–151. <https://doi.org/10.1186/s12874-018-0586-7>
- Visalakshi, J., & Jeyaseelan, L. (2014). Confidence interval for skewed distribution in outcome of change or difference between methods. *Clinical Epidemiology and Global Health*, 2(3), 117–120. <https://doi.org/10.1016/j.cegh.2013.07.006>
- Weisburd, D., & Britt, C. (2007). *Measures of association for nominal and ordinal variables* (pp. 335–380). Springer US. [https://doi.org/10.1007/978-0-387-34113-2\\_13](https://doi.org/10.1007/978-0-387-34113-2_13)
- Yu, H., Jiang, S., & Land, K. C. (2015). Multicollinearity in hierarchical linear models. *Social Science Research*, 53, 118–136. <https://doi.org/10.1016/j.ssresearch.2015.04.008>
- Yuan, K.-H., & Maxwell, S. E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30(2), 141–167. [https://articles.viriy.net/on\\_the\\_post\\_hoc\\_power\\_in\\_testing\\_mean\\_differences.pdf](https://articles.viriy.net/on_the_post_hoc_power_in_testing_mean_differences.pdf)
- Zhang, Y., Hedo, R., Rivera, A., Rull, R., Richardson, S., & Tu, X. M. (2019). Post hoc power analysis: Is it an informative and meaningful analysis? *General Psychiatry*, 32(4), e100069–e100069. <https://doi.org/10.1136/gpsych-2019-100069>

# Appendix A

## Common Statistical Symbols

Symbols are listed in rough alphabetical order. Some statistics are represented by more than one symbol, so some statistics are given more than once here.

More symbols are given in the [APA Manual](#).

Table A.1: Common Statistical Symbols

Symbol	Symbol Name/ Pronunciation	Meaning/Use/Interpretation
$\alpha$	Alpha	The probability of a false positive (a Type 1 error). Specifically, the probability of finding a given effect, value, etc. when the null hypothesis is true
$b$	$b$ -weight	The <i>unstandardized</i> coefficient of a term in a model, such as a general linear regression. In other words, the change in the outcome variable in terms of the units of the outcome and the given predictor/term.
$\beta$	Beta	Either: 1. The probability of a false negative (a Type 2 error) or 2. The standardized coefficient of a term in a model, such as a general linear regression
$1 - \beta$	One minus beta; power	The power of a test, statistic, etc. The probability of <i>not</i> making a false negative, i.e., the chance of not missing a real effect
$d$	Cohen's $d$	A measure of <b>effect size</b> for the difference between two means. It is computed simply as the difference in the means divided by the standard deviation of the means: Cohen's $d = \frac{\text{Mean}_1 - \text{Mean}_2}{SD}$ ; it is therefore a standardized measure that allows one to compare the size of effects across conditions, analyses, studies, etc.

Symbol	Symbol Name/ Pronunciation	Meaning/Use/Interpretation
$df$	Degrees of freedom	Roughly <sup>1</sup> , the amount (and source) of information in a set of data and related analyses. Each degree of freedom can be used—“spent”—to estimate some value in or about the data. For example, computing the mean of the data uses one degree of freedom; some statistics/analyses take up more than one degree of freedom to compute/conduct, and often using more degrees of freedom to conduct an analysis results in more powerful insights <sup>2</sup> .
$\eta^2$	Eta-squared	A measure of <b>effect size</b> for terms in a model—often a general linear regression. It is equivalent to $f^2$
$F$	$F$ -score	A statistic representing a value on a distribution (of the same name), used commonly to test for differences between groups/levels in an ANOVA. Conceptually, it represents the ratio of the size of an effect (e.g., a difference in means) to the amount of “error” in our measurement of the effect: i.e., a signal-to-noise ratio
$f^2$		The <b>effect size</b> of a term in a model, usually a general (or generalized) linear model.
$H_0$	Null hypothesis	The hypothesis to be refuted, it is usually set to be either that there is no difference between groups (e.g., there is no difference in outcomes between an experimental and a control group) or that there is no association between two variables (e.g., that two variables are not correlated). Abelson (1995) presents an accessible explanation of the role of the null hypothesis starting on <a href="#">page 8</a> .
$H_1$ or $H_A$	Alternative hypothesis	The hypothesis that there is, e.g., a difference between groups or an effect of an intervention/manipulation. It is often the proposed outcome of a study to be tested
MAD	Median absolute difference	A median distance of scores from the median. It is analogous to the standard deviation for a mean
$\mu$	Mu; mean	The mean of a population
$n$ or $N$		The number (count) of a sample, number of measurements, etc. Sometimes capital $N$ is used to denote the whole sample (or even sometimes for a population) while lower-case $n$ is used for a sub-sample / subgroup of a whole sample.

<sup>1</sup>More specifically, the number of values that can be computed in a given information space. If I have an equation like  $x + 2 = 3$ , then I have only one degree of freedom:  $x$  can only have one possible value (here, a 1 for the mathematically disinclined). If I instead have an equation like  $x + y = 3$ , there are two values to compute, and so I have two degrees of freedom. Importantly, notice that once I know one value—once I “spend” one degree of freedom, I then can determine the other value: If  $x = 1$ , then we can compute that  $y = 2$ ; if  $x = 0$ , then we can compute that  $y = 3$ . Therefore, “spending” degrees of freedom to estimate values (e.g., to compute the sample mean) uses up some of the total information available, but also lets us more accurately determine the value of other estimates, such as how far away a given person’s score is from that mean. We can use a data set’s degrees of freedom to make a certain number of insights, but the total number of insights is strictly limited by the size of our dataset. We could use those degrees of freedom to make different insights, but not an infinite number of insights.

<sup>2</sup>Devoting more of the available information to making a decision makes that decision more insightful. However, since any given set of data only has so many degrees of freedom—only so much information useful for making decisions—we should economize how much information we use to make a given decision.

Symbol	Symbol Name/ Pronunciation	Meaning/Use/Interpretation
$p$	<i>p-value</i>	<p>A probability; very often the probability of a false positive (a Type 1 error). More specifically, it indicates the probability that the researchers would have found the results that they did <i>even though there was no real difference</i>.</p> <p>More loosely and generally, saying, e.g., “we found something significant (<math>p &lt; .05</math>)” is taken to mean that the researchers believe there is less than a 5% chance that their results are “false positives.”</p>
$r$	Pearson correlation	A correlation between two continuous (interval/ratio) variables. It is also a measure of effect size; Cohen (1988) <i>suggests</i> that a correlation of .1 is “small,” that .3 is “medium,” and .5 is “large.”
$r_{pb}$	Point-biserial correlation	A correlation between a dichotomous variable and a continuous variable. The computation is equivalent to that for a Pearson correlation, so point-biserial correlations can be computed along with (mixed in with) Pearson correlations
$R^2$	R-squared	A measure of the total proportion of variance in the data that a given model can account for. It is always a measure for an entire model (i.e., all of the terms in a mode, including interaction terms and the intercept, if included). Like $r^2$ , it represents the variance accounted for. However, it is usually computed differently and can thus even take on negative values in <i>extreme cases</i> —usually for extremely-poorly fitting models. It is similar to <i>information criteria</i> , like AIC and BIC.
$\rho$	<i>Spearman's rho</i>	A correlation of the ranks of two ordinal (or even continuous) variables. It measures how much the ranking (from largest to smallest) of the values of one variable match the rankings of the values on an other variable. It is more robust than Pearson's $r$ and can be used for non-linear variables
$\sigma^2$	Sigma-squared; variance	Variance of a population
$S^2$	Variance	Variance of a sample
$\Sigma$	Sum	The sum (adding up) of a series of values, often the sum of all of the values of a given variable. The mean, for example, is the sum of scores divided by the number of scores, for example, the mean of variable $X$ is: $\bar{X} = \frac{\sum X}{N}$
$\sigma$	Sigma; standard deviation	Standard deviation of a population
$SD$	Standard deviation	Standard deviation of a sample

Symbol	Symbol Name/ Pronunciation	Meaning/Use/Interpretation
$t$	$t$ -score	A statistic representing a value on a distribution (of the same name), used commonly to test for differences between two (and only two) groups/levels. Like an $F$ -score, it represents the ratio of the size of an effect (e.g., a difference in means) to the amount of “error” in our measurement of the effect: i.e., a signal-to-noise ratio. In fact, it is simply the square root of a $F$ -score (i.e., $\sqrt{F} = t$ or, equivalently, $t^2 = F$ )
$\tau$	Kendall’s tau	A correlation between two ordinal variables. Like Spearman’s $\rho$ , it measures the ranks, but only whether the placement of participants, etc. are in the same ranks on both variables (for example, of two NPs both rank the severity of scoliosis of a set of patients in the same order)
$X$		$X$ is often used to represent a variable, often a dependent (input or predictor) variable
$\bar{X}$	Mean of $X$ ; $X$ overbar	The mean of variable $X$
$x$ -axis		The horizontal axis on a chart
$\text{var}(X)$	Variance	Variance of a variable (here—and often—a variable is represented by $X$ )
$\chi^2$	Chi-squared / Chi-square	A statistic representing a value on a distribution (of the same name) that resembles a normal distribution. It is very commonly used to test significance (e.g., if the frequency of events differs between two populations—like if the prevalence of COPD varies between genders)
$\bar{x}$	Mean (average)	A measure of “central tendency.” The mean is the sum of the scores divided by the number of scores
$Y$		$Y$ is often used to represent an outcome variable (e.g., an independent variable)
$y$ -axis		The vertical axis on a chart
$z$	$z$ -score	A standardized score, nearly always meaning standardized so that the (sample) mean is set to zero and the (sample) standard deviation is set to 1; this allows for direct comparisons between variables that have very different raw scores (e.g., comparing A1C levels to BMIs)

## Appendix B

# Common/Confusing Statistical & Scientific Terms

In addition to obtuse jargon<sup>1</sup>, both science in general and statistics in particular use fairly-common words in particular ways. Some of these common and confusing terms are listed below.

### B.1 Common/Confusing Statistical & Scientific Terms

Table B.1: Common/Confusing Statistical & Scientific Terms

Term	Meaning in Science & Statistics
Criterion	<p>The output/outcome variable, the variable that is measured to see the effects of other variables on it. (See Predictor, below).</p> <p>Also called:</p> <ul style="list-style-type: none"><li>• Dependent variable (DV)</li><li>• <a href="#">Exogenous variable</a></li><li>• Outcome (or “output”) variable</li><li>• Response variable</li><li>• Regressand</li><li>• Target</li></ul> <p>Calling it a criterion implies that it is being used as the basis or standard by which to test the importance of the predictors (or input variables) or even the success of our endeavors.</p>
Descriptive	<p>Descriptive statistics simply, well, describe a sample of data. They nearly always make no assumptions about the data—and make none about the population from which they were drawn (except perhaps that each datum is drawn independently from any other data and from the identical population.)</p>

<sup>1</sup>[“Generalized autoregressive conditional heteroskedasticity model”](#) anyone?

Term	Meaning in Science & Statistics
Factor	<p>Either:</p> <ul style="list-style-type: none"> <li>• A predictor variable of any type</li> <li>• An independent variable <b>manipulated or controlled</b> by researchers; in this sense, a factor is usually—but not necessarily—a categorical variable</li> <li>• A common source of variance/information in one or more ostensible variables</li> </ul>
Fixed	<p>No, nothing in science is every fixed (or broken : ). “Fixed” has two, mostly-different uses in statistics:</p> <ol style="list-style-type: none"> <li>1. A <b>fixed variable</b> is one in which all possible levels are present in the sample data. For example, if a sample of nursing home residents indicated that some had fallen while others had not (and there are no other possible categories), then falls would be a fixed factor. Variables in which not all levels are present in the data are called <b>random</b>. Often (but not always), these are continuous variables where it is impossible for all levels to be present in a sample of data, such as height which could be measured to nearly infinite levels. The important difference between fixed and random factors is that analyses must estimate the effects of levels <i>not</i> present for random factors, an issue that can eat up degrees of freedom, etc.</li> <li>2. <b>Fixed effects</b> in a linear regression (like in SPSS’s MIXED function, Chapter 11) are terms that have the same coefficient (e.g., same <math>\beta</math> weight for that term in the model) for all participants. For example, if all participants were assigned to either the Experimental or Control group, regardless of which hospital they were admitted to, then this would be a fixed effect. In hierarchical models, these are usually the levels that have something else nested in them, e.g., if patients were nested in hospitals. In this sense, <b>random</b> effects are terms that are nested within an other level<sup>2</sup>.</li> </ol>
iid	<p>An abbreviation that describes two, important characteristics of a set of data collected on a given variable. It stands for “independently and identically distributed,” meaning that:</p> <ol style="list-style-type: none"> <li>1. the value of each data point in a given variable is independent from the value of all/any other data point for that variable and</li> <li>2. each of those data points in that variable are drawn from the same distribution, e.g., they’re all drawn from a normal distribution.</li> </ol> <p>That the data in a given variable is iid is one of the most important assumptions in inferential statistics. Often, it can’t be violated without losing the validity of conclusions drawn from those data.</p>
Indicator	<p>A synonym for a dummy variable that indicates whether something is present or not present, e.g., recovered versus not recovered.</p>
Inferential	<p>Inferential analyses rely on assumptions being made about the population from which those data were drawn. This often includes assuming that the population is normally distributed. They are typically distinguished from descriptive statistics that make no (or fewer) assumptions about the population.</p>

Term	Meaning in Science & Statistics
Mean	<p>The average of a set of data. I'm including it in this list of common/confusing terms simply to note the main ways a mean can be computed, and their respective uses:</p> <ul style="list-style-type: none"> <li>• <b>Arithmetic</b> mean: The one you know, in which values are summed and then divided by the number of values. It is used when there are no particular reasons to use an other method.</li> <li>• <b>Geometric</b> mean: Values are multiplied instead of added. We then <a href="#">take the <math>n</math>th root of this product</a>. Geometric means are useful to compare very different values; to get the mean of percents, proportions, etc., and when the values are related to each other (like all the percents being of the same thing, like inflation).</li> <li>• <b>Harmonic</b> mean: It is computed as "<a href="#">the reciprocal of the average of the reciprocals of the data values</a>". It is used when we want to reduce the weight of larger values, such as when a distribution is positively skewed (i.e., has disproportionate number of large values). An example is length of time where events can't be shorter than zero, but sometimes can take a lot longer than they should<sup>3</sup></li> <li>• <b>Weighted</b> mean: Any of the above types of means can also be weighted. In a weighted mean, some of the values are given heavier<sup>4</sup> weights (their values are multiplied by some number to make them affect the overall mean differently) than other values so that those weighted values contribute more to the overall mean. This is commonly done when we were unable to sample enough people of a certain type, such as when we were unable to sample enough members of a minoritized group.</li> </ul>
Multiple vs. multivariate (e.g., multiple vs. multivariate linear regression)	<ul style="list-style-type: none"> <li>• <b>"Multiple"</b> indicates that there is more than one linear regression equation; this would happen if there is <b>more than one outcome</b></li> <li>• <b>"Multivariate"</b> indicates that there is <b>more than one predictor</b>.</li> </ul>
Non-ostensible	<p>A variable that cannot be directly observed or other perceived. These are usually theoretical and abstract concepts—constructed ideas—that are assumed to give rise to "ostensible" variables that can be empirically perceived. Other terms for these and related concepts are:</p> <ul style="list-style-type: none"> <li>• Latent</li> <li>• Unobserved</li> </ul> <p>These are similar—sometimes even defined by—the factors that emerge from factor (latent variable) analysis.</p>
Non-parametric	<p>Non-parametric analyses make no (or few) assumptions about the population distribution, viz., that it is normally distributed. Non-parametric analyses tend to be more robust than parametric analyses; they also tend to be used for variables of lower measurement levels (e.g., for ordinal instead of interval/ratio data).</p>
Ostensible	<p>A variable that can be directly, empirically observed. This is used to distinguish a variable from non-ostensible ones that are theorized to manifest in observable ways (through ostensible variables). Other terms for these and related concepts are:</p> <ul style="list-style-type: none"> <li>• <a href="#">Manifest variable</a></li> <li>• Observable variable</li> </ul>

Term	Meaning in Science & Statistics
Parametric	Parametric analyses are inferential analyses that make assumptions about the mathematical values (“parameters”) about the population’s distribution. Nearly always, this is the assumption that the population distribution is normal. Therefore, parametric analyses are those that assume normality. The term is also nearly always used to contrast these analyses with non-parametric ones that do not require (as many) assumptions about the population distribution. Making fewer assumptions, non-parametric analyses tend to be more robust.
Parsimony	The desirable trait of communicating efficiently, saying a lot of information clearly and succinctly. It is also said of explanations and theories, suggesting a strong explanation that is “elegantly” simple yet generally useful.
Population	A well-defined and -delimited group of individuals (patients, nurses, etc.) about which insights are made based on a smaller sample of members of that population. The sample are those chosen to be studied directly; the population are those to whom conclusions made from the sample can be justifiably applied.
Power	The probability to detect a real effect. The chance <i>not</i> to make a false negative (Type 2) error.
Predict	<p>In statistics, prediction is our ability to use what we know to make inferences about what we don’t. This could be information we have from the past and present that we use to guess at the future. But it could also be using known information about the past to make inferences about other, past events we don’t know about. And yes, this certainly includes using sample data to infer population values.</p> <p>A rather clever use of prediction is to randomly split a larger set of data in half. Use half to create a model with a given set of parameters &amp; values. And then see how well those parameters, etc. predict the other half of the data. This allow us to conduct a very authentic test of how good our estimates were.</p>
Predictor	<p>One of the many terms used to indicate the variables added to a linear regression model to test the effect on the outcome.</p> <p>Also called:</p> <ul style="list-style-type: none"> <li>• Explanatory variable</li> <li>• Independent variable (IV)</li> <li>• Input variable</li> <li>• Regressor</li> </ul>
Random	<p>Generally, the crux of randomness is that the value is unbiased and—in the long run—therefore an accurate representation of the true state. However, it can also refere to:</p> <ul style="list-style-type: none"> <li>• The process of selecting a participant, level, etc. without bias so that any value is either equally likely to be chosen or at least chosen by the same rules &amp; odds as any other value</li> <li>• A “random variable” is a rather generic term for any empirical value that can take multiple values, and that the value is takes is “iid”: independent from and identically distributed as all other measurements taken on that variable</li> <li>• A “random effect” is a term in a hierarchical linear modle (aka multilevel model, Chapter 11) that is nested within an other variable, like patients nested in hospitals.</li> </ul>

Term	Meaning in Science & Statistics
Spell	In longitudinal analyses, an occasion on which some outcome is present. For example, each time when a woman is pregnant or when someone with a substance abuse disorder recidivates. “Spell” is usually (but not always) used when an event can happen more than once in a longitudinal study.
Variance	A measure of dispersion. It is the square of the standard deviation (when computed for the dispersion of a variable); it is the square of the distance from a regression line (when computed in a linear regression). It is also a measure of the total amount of information in a variable; the more information, the richer a variable is, but the more there is to try to understand.
Wave	An instance of data collection at a given point in time. This terms is usually only used when there are more than one data collection occasions, viz., when a study is longitudinal. Also called: <ul style="list-style-type: none"> <li>• Event</li> <li>• Endogenous variable</li> <li>• Instance</li> <li>• Period</li> <li>• Phase</li> <li>• Time point</li> </ul>

<sup>2</sup>More specifically, these are terms where the intercept and/or slope is allowed to vary for each level, e.g., for each patient.

<sup>3</sup>Which, of course, never applies to dissertations.

<sup>4</sup>Sure, you could also/instead give lighter weights to some values—e.g., values from over-represented groups—but we usually instead give heavier weights to members of *underrepresented* groups.

## B.2 Terms for Different Types of Analyses

Table B.2: Terms for Different Types of Analyses

Analysis	Measurement Level		Uses & Notes
	Outcome Variables(s)	Predictor(s)	
ANOVA	Continuous	Nominal	<ul style="list-style-type: none"> <li>• Understood by many, so easily communicated</li> <li>• Variance determined by ordinary least squares</li> <li>• Typically only used to test significance of individual (main effect and/or interaction) terms</li> </ul>
One-way ANOVA etc.	Continuous	One Nominal	<ul style="list-style-type: none"> <li>• The “one-way” indicates that there is only one predictor.</li> <li>• If there are two predictors, it’s instead called a “two-way” ANOVA.</li> <li>• We <i>could</i> also use “three-way” ANOVA, etc., but we instead just give up and call them “multi-way” ANOVAs when there are <math>\geq 3</math> nominal predictors.</li> </ul>
ANCOVA	Continuous	$\geq 1$ Nominal & $\geq 1$ Continuous	<ul style="list-style-type: none"> <li>• Contains one or more continuous “covariates”</li> <li>• Variance determined by ordinary least squares</li> <li>• Typically only used to test significance of individual (main effect and/or interaction) terms</li> </ul>
Repeated-measures ANOVA	Continuous	Nominal	<ul style="list-style-type: none"> <li>• A repeated-measures ANOVA not only tests the effect of <math>\geq 1</math> predictors, but also whether the effect(s) of the predictor(s) changes over time.</li> <li>• The times when data are collected must be evenly spaced</li> <li>• An, e.g., “repeated-measures ANCOVA” is an ANCOVA (with <math>\geq 1</math> nominal and <math>\geq 1</math> continuous) that tests differences in predictors’ effects over time</li> </ul>
MANOVA	$\geq 2$ Continuous	Nominal	<ul style="list-style-type: none"> <li>• A MANOVA includes two or more outcome variables.</li> <li>• The benefit of conducting a MANOVA over two (or more), separate ANOVAs (one for each outcome) is that a MANOVA also tests for (and accounts for) the relationships between the outcomes</li> </ul>
General linear model	Continuous	Nominal and/or continuous	<p>“General linear model” is the term for any linear model that:</p> <ul style="list-style-type: none"> <li>• Has a continuous outcome</li> <li>• Assumes a linear relationship between the predictors and the outcome.</li> </ul> <p>General linear models include ANOVAs (and their ilk), <i>t</i>-tests, <i>F</i>-tests, etc. They are one of the types of generalized linear models.</p>

Measurement Level			
Generalized linear model	Any	Any	<ul style="list-style-type: none"> <li>• “Generalized linear model” is the term used for <i>any</i> inferential analysis that uses the general formula of <math>Y = b_0 + b_1X_1 + \dots + b_kX_k + e</math> to describe the relationship between <math>k</math> predictors and <math>\geq 1</math> outcomes. This includes general linear models (and thus ANOVAs, etc.), logistic regression, structural equation models, etc.</li> <li>• Although the formula is written as a linear equation, generalized linear models can model many types of non-linear relationships between predictors and outcomes. <ul style="list-style-type: none"> <li>◦ They can test dichotomous outcomes (viz., logistic regression), logarithmic outcomes, etc.</li> </ul> </li> <li>• Data can be <a href="#">heteroskedastic</a>.</li> <li>• Variables need not be normally distributed (but their distribution must still be correctly modeled by the equation)</li> </ul> <p>The term “generalized linear model” is not as commonly used as “general linear model” (instead one uses the term for the type of model conducted), but it is still useful to know the difference.</p>
Logistic regression	Dichotomous	Nominal or continuous	Logistic regression models are used to estimate the effects on a dichotomous outcome, such as whether or not a patient has a given condition.
Ordinal regression	Ordinal	Nominal or continuous	Although I believe we can often treat ordinal variables as continuous (i.e., lump them in with interval & ratio variables), there are times when we should indeed treat a variable as assuming nothing more than ordinalism, such as when we have only a few levels and know the distances between them vary, but don’t know how much.
Multinomial (logistic) regression	Nominal	Nominal or continuous	These are very similar to logistic regression, differing mainly in that there are more than two levels to the outcome variable. Like logistic regression, one of the main goals of multinomial regression is to determine (predict) the category one falls into based on one’s values for the predictors.
Multiple linear regression	Continuous	$\geq 2$ Nominal or continuous	This is a general term for a linear model that has two or more predictors.

<b>Measurement Level</b>			
<i>t</i> -test	Continuous	One dichotomous	<ul style="list-style-type: none"> <li>• <i>t</i>-Tests are used to test the difference in two values. They are used, for example, to test:</li> <li>• The difference between the means of two groups               <ul style="list-style-type: none"> <li>◦ Such as two study groups (e.g., experimental &amp; control),</li> <li>◦ Or the means for each level of a nominal variable with two levels (e.g., those diagnosed / not diagnosed with a condition)</li> </ul> </li> <li>• Changes in one outcome at two, different points in time (a “paired” <i>t</i>-test).</li> <li>• Whether a single mean is different than some value, e.g., if the mean is not zero (a “one-sample <i>t</i>-test”). These are often done in linear regressions to test if a given parameter weight is significantly different from zero.</li> </ul>

## Appendix C

# Statistical Analysis Decision Trees and Guides

One of my goals for this curriculum is to empower you to be able to use models—especially generalized linear models—flexibly. To think of the sorts of research questions you want answered, how to operationalize those questions, and then design analyses around those questions. One simple (or simplistic) way to approach that is to think of what level of measurement your variables represent and then choose the analysis recommended for them.

That simpler approach is simply addressed through a statistical analysis decision tree or dedicated guide. This appendix presents that. Or rather, presents “those” since this appendix is—and may long remain—little more than a collection of links to others who have done this already. Even though I’ve already curated and culled from what’s out there, this appendix is still less of a venerable decision tree and more a brambly new-growth forest. I will nonetheless try to forge a path through these decision trees best I can.

Table B.2 presents a complementary reference by describing some common analyses and the types of outcome and predictor variables used in each.

One final note: As you might expect given the varied and ad hoc nature of naming in statistics, “decision tree” *also* denotes a [type of analysis](#) (actually an analytic strategy) that are beyond the pale of this curriculum.

## C.1 References and Guides

The following are sources that discuss guidelines, etc. for which statistic to choose.

### C.1.1 Correlations & Associations

- [Khamis \(2008\)](#) clearly presents which measure of association/correlation to use with various types of data, along with some guides on interpreting the strengths of these measures. Their recommendations are summarized in this table, reproduced from their [Summary](#) section:

Table C.1: Types of Correlation Statistics

Variable Y	Variable X		
	Nominal	Ordinal	Continuous
Nominal	$\phi$ coefficient or Goodman & Kruskal's $\lambda$	Rank biserial <sup>1</sup>	Point biserial
Ordinal	Rank biserial	Kendall's $\tau_b$ or Spearman's $\rho$	Kendall's $\tau_b$ or Spearman's $\rho$
Continuous	Point biserial	Kendall's $\tau_b$ or Spearman's $\rho$	Pearson's $r$ or Spearman's $\rho$

## C.2 Simple Graphics

The following trees are simple files that organize the analyses one commonly uses to test straight-forward hypothesis tests between relatively small groups of variables, e.g., one outcome and one or two predictors. They all top out at ANOVAs, and thus effectively fill in the gap I left unfilled for what is covered before the model building focused in our curriculum.

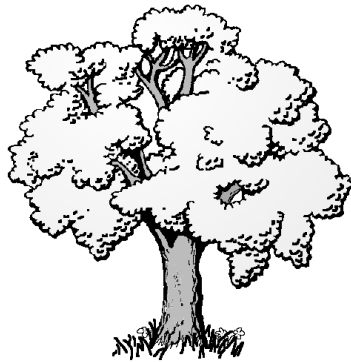
- [Howell \(2008\)](#) covers analyses from correlations to ANOVAs. The benefit of this tree is its simplicity; the deficit is its lack of specificity between parametric and non-parametric analyses.
- [Corston & Colman's \(2000\)](#) tree is also a simple “cheat sheet” file like Howell's, but contains more information about distinguishing between parametric and non-parametric tests (via the level of measurement of the given variables).

## C.3 Online Trees

These are websites that let you choose the analysis by answering a series of questions. They tend to be more thorough than the simple graphics, but require a more involved process to get to the solution.

- [MicrOsiris](#)'s decision tree allows one to step through questions to determine what analysis to conduct; it also provides a nice [summary page](#) that indicates which function to use to conduct a given analysis in SPSS, SAS, and their own freeware stat program, [MicrOsiris](#).
- [NIST's Decision Tree for Key Comparisons](#) is more than just that. As the [About](#) page says, the tree “guides users through a series of hypothesis tests intended to help them in deciding upon an appropriate statistical model for their particular data.” One can first enter or upload (a .csv file) to the site and then see what their tree recommends as analyses for those data. Pretty cool, huh?
- [Statistical Test Flowchart](#) doesn't take as many steps as, e.g., MicrOsis's tree. It thus presents less specific results at the end, but it gives more of an explanation of what it recommends along with links to how to conduct the given analysis in R, SPSS, and Stata.

<sup>1</sup>Weisburd and Britt (2007) give a good, further coverage of analyzing associations between nominal and ordinal data.





# Index

- ANOVA Family, 125, 230, 292
- Applying a new style to an existing LO Writer document, 205
- Applying a new style to an existing MS Word document, 207
- Central Limit Theorem, 9
- Chi-Squared ( $\chi^2$ ), 117
- Cohen's d, 72
- Cohen's f, 72
- Correlations, 61
- Correlations as Effect Size Statistics, 75
- Creating Error Bars in SPSS, 32
- Crosstables, 115
- Degrees of Freedom, 35
- Descriptive Statistics in SPSS, 23
- Dummy Variables, 21, 115, 133
- Dummy Variables in SPSS, 268
- Effect Size, 37, 71
- Effect Size Conversions, 83
- Effect Size Statistics for ANOVAs, etc., 72
- Effect Size Statistics for Mean Differences, 72
- Effect Size Statistics for Odds and Risk Ratios, 77
- Effect Size Statistics for Terms in Linear Regression Models, 75
- Effect Size, explanation of "small," "medium," and "large" effects, 78
- Epsilon-squared, 77
- Eta-squared, 75
- Excel Formulas, 223
- Excel Shortcuts, 222
- F-Test, 37
- Figures and Graphs, 120, 132
- Freeze Rows or Columns in Excel Formulas, 227
- Information Criteria, 157
- Linear Regression Models, 55, 129
- Long-Format Data, 147
- Longitudinal Analyses, 137
- Mean Square, 36
- Merging Data, 149
- Modifying Charts in SPSS, 123
- Multicollinearity, 103
- Omega-squared, 76
- Partial Correlations, 61
- Principal Component Analysis, 7
- Q-Q Plots, 10
- Reference Managers, 212
- Signal-to-Noise Ratio, 54
- Source Tables, 35, 100, 101, 128
- Sphericity, 141
- SPSS Data Editor, 260
- SPSS Global Options, 17, 259
- SPSS Measurement Scales, 19
- SPSS Output Window, 263
- SPSS Syntax, 150, 264
- SPSS Value Labels, 19
- SPSS Variable View, 261
- Standardized Variables, 19
- Standardizing Variables in SPSS, 274
- Stepwise Regression, 98
- Style Templates, 203
- t-Test, 229
- Transposing Data, 222
- Variance and Covariance, 59
- Variance Inflation Factor (VIF), 103
- Wide-Format Data, 147
- Writing Results, 94, 184
- Zotero, 214