

# ReCentering Psych Stats: Multivariate Modeling

Lynette H. Bikos, PhD, ABPP

25 Sep 2023



# Contents

<b>1 Scrubbing</b>	<b>23</b>
1.1 Navigating this Lesson . . . . .	23
1.1.1 Learning Objectives . . . . .	23
1.1.2 Planning for Practice . . . . .	24
1.1.3 Readings & Resources . . . . .	24
1.1.4 Packages . . . . .	24
1.2 Workflow for Scrubbing . . . . .	24
1.3 Research Vignette . . . . .	26
1.4 Working the Problem . . . . .	26
1.4.1 intRavenous Qualtrics . . . . .	26
1.4.1.1 Option 1. Upload an .rds file . . . . .	28
1.4.1.2 Option 2. Upload a .csv file . . . . .	29
1.4.2 About the <i>Rate-a-Recent-Course Survey</i> . . . . .	29
1.4.3 The Codebook . . . . .	30
1.5 Scrubbing . . . . .	31
1.5.1 Tools for Data Manipulation . . . . .	31
1.5.2 Inclusion and Exclusion Criteria . . . . .	32
1.5.3 Renaming Variables . . . . .	35
1.5.4 Downsizing the Dataframe . . . . .	36
1.6 Toward the APA Style Write-up . . . . .	37
1.6.1 Method/Procedure . . . . .	37
1.7 Practice Problems . . . . .	37
1.7.1 Problem #1: Rework the Chapter Problem . . . . .	37
1.7.2 Problem #2: Use the <i>Rate-a-Recent-Course Survey</i> , Choosing Different Variables . . . . .	37
1.7.3 Problem #3: Other data . . . . .	38

1.7.4	Grading Rubric . . . . .	38
1.8	Bonus Track: . . . . .	38
1.8.1	Importing data from an exported Qualtrics .csv file . . . . .	38
<b>2</b>	<b>Scoring</b>	<b>41</b>
2.1	Navigating this Lesson . . . . .	41
2.1.1	Learning Objectives . . . . .	41
2.1.2	Planning for Practice . . . . .	41
2.1.3	Readings & Resources . . . . .	42
2.1.4	Packages . . . . .	42
2.2	Workflow for Scoring . . . . .	43
2.3	Research Vignette . . . . .	43
2.4	On Missing Data . . . . .	44
2.4.1	Data Loss Mechanisms . . . . .	44
2.4.2	Diagnosing Missing Data Mechanisms . . . . .	45
2.4.3	Managing Missing Data . . . . .	46
2.4.4	Available Information Analysis (AIA) . . . . .	47
2.5	Working the Problem . . . . .	48
2.5.1	Variable Planning and Preparation . . . . .	48
2.5.2	Missing Data Analysis: Whole df and Item level . . . . .	54
2.5.3	Analyzing Missing Data Patterns . . . . .	56
2.5.4	Can we identify the missing mechanisms? . . . . .	57
2.6	Scoring . . . . .	57
2.6.1	Reverse scoring . . . . .	57
2.7	Missing Analysis: Scale level . . . . .	59
2.8	Revisiting Missing Analysis at the Scale Level . . . . .	61
2.8.1	Scale Level: Patterns of Missing Data . . . . .	62
2.8.2	R-ready for Analysis . . . . .	63
2.9	The APA Style Write-Up . . . . .	64
2.10	Results . . . . .	64
2.11	Practice Problems . . . . .	65
2.11.1	Problem #1: Reworking the Chapter Problem . . . . .	65
2.11.2	Problem #2: Use the <i>Rate-a-Recent-Course Survey</i> , Choosing Different Variables . . . . .	65
2.11.3	Problem #3: Other data . . . . .	65
2.11.4	Grading Rubric . . . . .	65

CONTENTS	5
----------	---

<b>3 Data Dx</b>	<b>67</b>
3.1 Navigating this Lesson . . . . .	67
3.1.1 Learning Objectives . . . . .	67
3.1.2 Planning for Practice . . . . .	68
3.1.3 Readings & Resources . . . . .	68
3.1.4 Packages . . . . .	68
3.2 Workflow for Preliminary Data Diagnostics . . . . .	68
3.3 Research Vignette . . . . .	69
3.4 Internal Consistency of Scales/Subscales . . . . .	71
3.5 Distributional Characteristics of the Variables . . . . .	75
3.5.1 Evaluating Univariate Normality . . . . .	75
3.5.2 Pairs Panels . . . . .	78
3.6 Evaluating Multivariate Normality . . . . .	80
3.7 A Few Words on Transformations . . . . .	83
3.8 The APA Style Write-Up . . . . .	83
3.8.1 Data Diagnostics . . . . .	83
3.9 A Quick Regression of our Research Vignette . . . . .	85
3.9.1 Results . . . . .	85
3.10 Practice Problems . . . . .	87
3.10.1 Problem #1: Reworking the Chapter Problem . . . . .	87
3.10.2 Problem #2: Use the <i>Rate-a-Recent-Course</i> Survey, Choosing Different Variables . . . . .	88
3.10.3 Problem #3: Other data . . . . .	88
3.10.4 Grading Rubric . . . . .	88
3.11 Homeworked Example . . . . .	88
3.11.1 Scrubbing . . . . .	89
3.11.2 Scoring . . . . .	91
3.11.3 Data Dx . . . . .	99
3.11.4 Results . . . . .	107
<b>4 Multiple Imputation (A Brief Demo)</b>	<b>111</b>
4.1 Navigating this Lesson . . . . .	111
4.1.1 Learning Objectives . . . . .	111
4.1.2 Planning for Practice . . . . .	112

4.1.3	Readings & Resources . . . . .	112
4.1.4	Packages . . . . .	112
4.2	Workflow for Multiple Imputation . . . . .	113
4.3	Research Vignette . . . . .	113
4.4	Multiple Imputation – a Super Brief Review . . . . .	115
4.4.1	Steps in Multiple Imputation . . . . .	116
4.4.2	Statistical Approaches to Multiple Imputation . . . . .	117
4.5	Working the Problem . . . . .	117
4.5.1	Selecting and Formatting Variables . . . . .	118
4.5.2	Creating Composite Variables . . . . .	120
4.5.3	The Multiple Imputation . . . . .	122
4.5.4	Creating Scale Scores . . . . .	124
4.6	Multiple Regression with Multiply Imputed Data . . . . .	125
4.7	Toward the APA Style Write-up . . . . .	127
4.7.1	Method/Data Diagnostics . . . . .	127
4.7.2	Results . . . . .	128
4.8	Multiple imputation considerations . . . . .	128
4.9	Practice Problems . . . . .	129
4.9.1	Problem #1: Reworking the Chapter Problem . . . . .	129
4.9.2	Problem #2: Use the <i>Rate-a-Recent-Course</i> Survey, Choosing Different Variables . . . . .	129
4.9.3	Problem #3: Other data . . . . .	129
4.9.4	Grading Rubric . . . . .	130
4.10	Homeworked Example . . . . .	130
4.10.1	Scrubbing . . . . .	130
4.10.1.1	Format any variables that shouldn't be imputed in their raw form .	132
4.10.1.2	Multiply impute a minimum of 5 sets of data . . . . .	133
4.10.1.3	Run a regression (for multiply imputed data) with at least three variables . . . . .	136
4.10.1.4	APA style write-up of the multiple imputation section of data diagnostics . . . . .	138
4.10.1.5	APA style write-up regression results . . . . .	139

<b>5 Simple Mediation</b>	<b>143</b>
5.1 Navigating this Lesson . . . . .	143
5.1.1 Learning Objectives . . . . .	143
5.1.2 Planning for Practice . . . . .	144
5.1.3 Readings & Resources . . . . .	144
5.1.4 Packages . . . . .	145
5.2 Estimating Indirect Effects (the analytic approach often termed <i>mediation</i> ) . . . . .	145
5.2.1 The definitional and conceptual . . . . .	145
5.3 Workflow for Simple Mediation . . . . .	147
5.4 Super Simple Mediation in <i>lavaan</i> : A focus on the mechanics . . . . .	150
5.4.1 Simulate Fake Data . . . . .	151
5.4.2 Specify Mediation Model . . . . .	151
5.4.3 Interpret the Output . . . . .	155
5.4.4 A Figure and Table . . . . .	157
5.4.5 Results . . . . .	161
5.5 Research Vignette . . . . .	161
5.5.1 Data Simulation . . . . .	162
5.5.2 Scrubbing, Scoring, and Data Diagnostics . . . . .	167
5.5.3 Specify the Model in <i>lavaan</i> . . . . .	169
5.5.4 Interpret the Output . . . . .	173
5.5.5 A Figure and a Table . . . . .	173
5.5.6 Results . . . . .	176
5.6 Considering Covariates . . . . .	176
5.6.1 A Figure and a Table . . . . .	180
5.6.2 APA Style Write-up . . . . .	183
5.7 STAY TUNED . . . . .	184
5.8 Residual and Related Questions... . . . . .	184
5.9 Practice Problems . . . . .	185
5.9.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed . . . . .	185
5.9.2 Problem #2: Rework the research vignette, but swap one or more variables .	185
5.9.3 Problem #3: Use other data that is available to you . . . . .	185
5.9.4 Grading Rubric . . . . .	186
5.10 Homeworked Example . . . . .	186

5.10.1 Assign each variable to the X, Y, or M roles (ok but not required to include a covariate) . . . . .	186
<b>6 Complex Mediation</b>	<b>197</b>
6.1 Navigating this Lesson . . . . .	197
6.1.1 Learning Objectives . . . . .	197
6.1.2 Planning for Practice . . . . .	198
6.1.3 Readings & Resources . . . . .	198
6.1.4 Packages . . . . .	198
6.2 Complex Mediation . . . . .	199
6.3 Workflow for Complex Mediation . . . . .	199
6.4 Parallel Mediation . . . . .	201
6.4.1 A Mechanical Example . . . . .	202
6.4.1.1 Data Simulation . . . . .	202
6.4.1.2 Specifying <i>lavaan</i> code . . . . .	203
6.4.1.3 A note on indirect effects and confidence intervals . . . . .	207
6.4.1.4 Figures and Tables . . . . .	207
6.4.1.5 APA Style Writeup . . . . .	211
6.4.2 Research Vignette . . . . .	212
6.4.2.1 Data Simulation . . . . .	213
6.4.3 Scrubbing, Scoring, and Data Diagnostics . . . . .	216
6.4.3.1 Specifying the <i>lavaan</i> model . . . . .	218
6.4.3.2 Table and Figure . . . . .	219
6.4.3.3 APA Style Writeup . . . . .	223
6.5 Serial Multiple Mediator Model . . . . .	225
6.5.1 We stick with the Lewis et al. [2017] example, but modify it. . . . .	226
6.5.2 Specify the <i>lavaan</i> model . . . . .	226
6.5.2.1 Table and Figure . . . . .	227
6.5.3 APA Style Writeup . . . . .	231
6.6 STAY TUNED . . . . .	232
6.7 Troubleshooting and FAQs . . . . .	232
6.8 Practice Problems . . . . .	233
6.8.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed . . . . .	234
6.8.2 Problem #2: Rework the research vignette, but swap one or more variables .	234

6.8.3	Problem #3: Use other data that is available to you . . . . .	234
6.8.4	Grading Rubric . . . . .	234
6.9	Homeworked Example . . . . .	234
<b>7</b>	<b>Simple Moderation in OLS and MLE</b>	<b>247</b>
7.1	Navigating this Lesson . . . . .	247
7.1.1	Learning Objectives . . . . .	247
7.1.2	Planning for Practice . . . . .	248
7.1.3	Readings & Resources . . . . .	248
7.1.4	Packages . . . . .	249
7.2	On <i>Modeling</i> : Introductory Comments on the simultaneously invisible and paradigm-shifting transition we are making . . . . .	251
7.2.1	NHST versus modeling . . . . .	251
7.2.2	Introducing: <i>The Model</i> . . . . .	252
7.3	OLS to ML for Estimation . . . . .	252
7.3.1	Ordinary least squares (OLS) . . . . .	252
7.3.2	Maximum likelihood estimation (MLE): A brief orientation . . . . .	254
7.3.3	OLS and MLE Comparison . . . . .	255
7.3.4	Hayes and PROCESS (aka conditional process analysis) . . . . .	255
7.4	Introducing the <i>lavaan</i> package . . . . .	256
7.4.1	The FIML magic for which we have been waiting . . . . .	256
7.5	Picking up with Moderation . . . . .	258
7.6	Workflow for a Simple Moderation . . . . .	260
7.7	Research Vignette . . . . .	260
7.7.1	Simulate Data from the Journal Article . . . . .	261
7.8	Working the Simple Moderation with OLS and MLE . . . . .	264
7.8.1	OLS with <i>lm()</i> . . . . .	264
7.8.1.1	An APA Style Write-up of OLS results . . . . .	267
7.8.2	MLE with <i>lavaan::sem()</i> . . . . .	268
7.8.3	Tabling the data . . . . .	274
7.8.4	APA Style Writeup . . . . .	274
7.9	Residual and Related Questions... . . . . .	275
7.10	Practice Problems . . . . .	275
7.10.1	Problem #1: Rework the research vignette as demonstrated, but change the random seed . . . . .	275

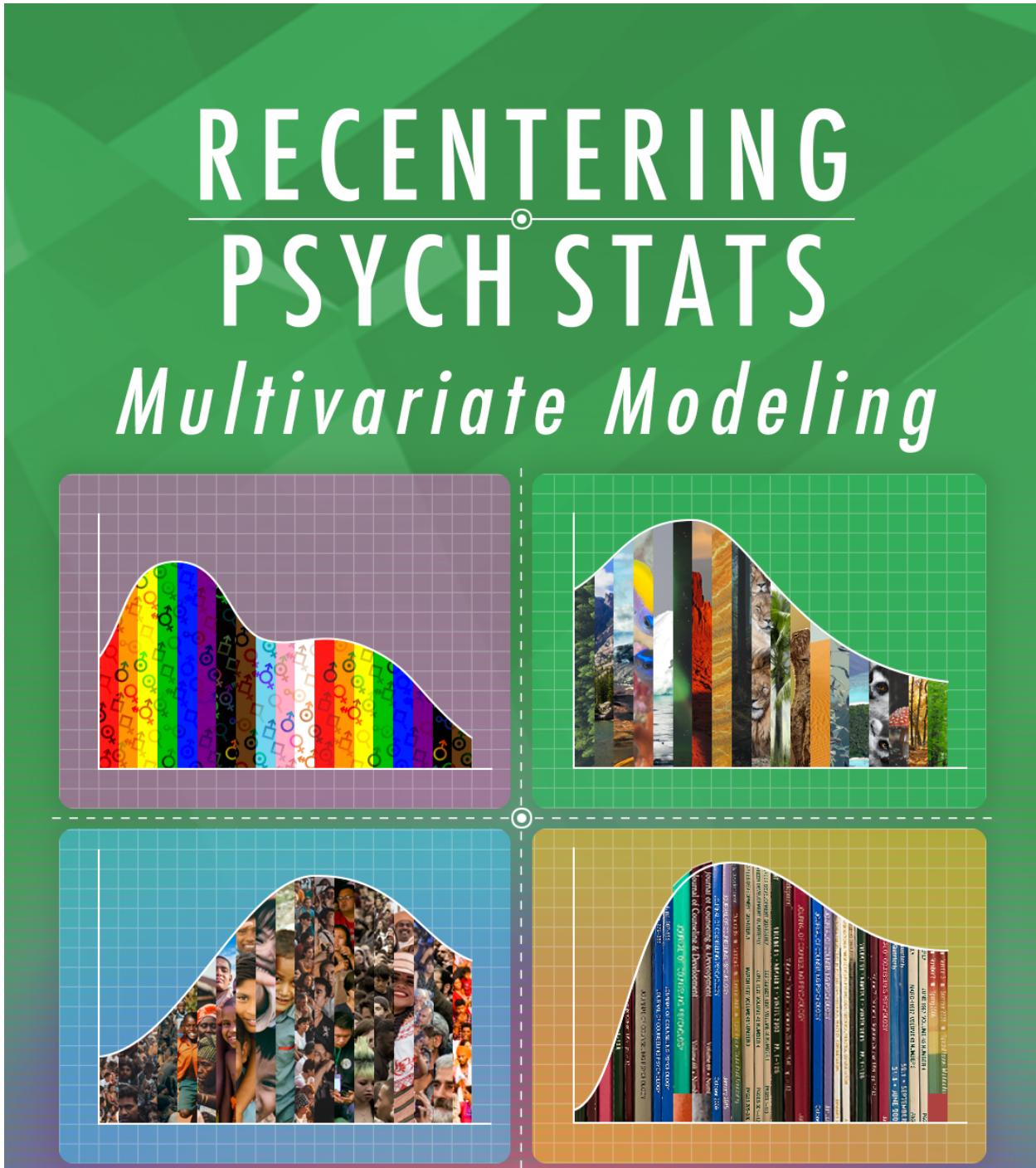
7.10.2 Problem #2: Rework the research vignette, but swap one or more variables . . . . .	276
7.10.3 Problem #3: Use other data that is available to you . . . . .	276
7.11 Bonus Track: . . . . .	277
<b>8 Moderated Mediation</b>	<b>281</b>
8.1 Navigating this Lesson . . . . .	281
8.1.1 Learning Objectives . . . . .	281
8.1.2 Planning for Practice . . . . .	282
8.1.3 Readings & Resources . . . . .	282
8.1.4 Packages . . . . .	283
8.2 Conditional Process Analysis . . . . .	284
8.2.1 The definitional and conceptual . . . . .	284
8.2.2 Hayes' [2018] Piecewise Approach to Building Models . . . . .	286
8.3 Workflow for Moderated Mediation . . . . .	287
8.4 Research Vignette . . . . .	287
8.4.1 Simulating the data from the journal article . . . . .	288
8.4.2 Quick peek at the data . . . . .	290
8.5 Working the Moderated Mediation . . . . .	292
8.5.1 Piecewise Assembly of the Moderated Mediation . . . . .	294
8.5.1.1 Analysis #1: A simple moderation . . . . .	294
8.5.1.2 Analysis #2: Another simple moderation . . . . .	299
8.5.1.3 Analysis #3: A simple mediation . . . . .	303
8.6 The Moderated Mediation: A Combined analysis . . . . .	307
8.6.1 Specification in <i>lavaan</i> . . . . .	307
8.6.2 A quick plot . . . . .	310
8.6.3 Beginning the interpretation . . . . .	311
8.6.4 Tabling the data . . . . .	312
8.6.4.1 Conditional Indirect effects . . . . .	313
8.6.4.2 Conditional Direct effect . . . . .	314
8.6.5 Model trimming . . . . .	314
8.6.6 APA Style Write-up . . . . .	314
8.7 Residual and Related Questions... . . . . .	316
8.8 Practice Problems . . . . .	317

8.8.1	Problem #1: Rework the research vignette as demonstrated, but change the random seed . . . . .	317
8.8.2	Problem #2: Rework the research vignette, but swap one or more variables .	318
8.8.3	Problem #3: Use other data that is available to you . . . . .	318
8.9	Bonus Track: . . . . .	319





# BOOK COVER



at the GitHub repository:

- Formatted as an [html book](#) via GitHub Pages available
- As a [PDF](#)
- As an [ebook](#)
- As a [Word](#)

All materials used in creating this OER are available at its [GitHub repo](#).



# PREFACE

If you are viewing this document, you should know that this is a book-in-progress. Early drafts are released for the purpose teaching my classes and gaining formative feedback from a host of stakeholders. The document was last updated on 25 Sep 2023. Emerging volumes on other statistics are posted on the [ReCentering Psych Stats](#) page at my research team's website.

## [Screencasted Lecture Link](#)

To *center* a variable in regression means to set its value at zero and interpret all other values in relation to this reference point. Regarding race and gender, researchers often center male and White at zero. Further, it is typical that research vignettes in statistics textbooks are similarly seated in a White, Western (frequently U.S.), heteronormative, framework. The purpose of this project is to create a set of open educational resources (OER) appropriate for doctoral and post-doctoral training that contribute to a socially responsive pedagogy – that is, it contributes to justice, equity, diversity, and inclusion.

Statistics training in doctoral programs are frequently taught with fee-for-use programs (e.g., SPSS/AMOS, SAS, MPlus) that may not be readily available to the post-doctoral professional. In recent years, there has been an increase and improvement in R packages (e.g., *psych*, *lavaan*) used for in analyses common to psychological research. Correspondingly, many graduate programs are transitioning to statistics training in R (free and open source). This is a challenge for post-doctoral psychologists who were trained with other software. This OER will offer statistics training with R and be freely available (specifically in a GitHub repository and posted through GitHub Pages) under a Creative Commons Attribution - Non Commercial - Share Alike license [CC BY-NC-SA 4.0].

Training models for doctoral programs in HSP are commonly scholar-practitioner, scientist-practitioner, or clinical-scientist. An emerging model, the *scientist-practitioner-advocacy* training model incorporates social justice advocacy so that graduates are equipped to recognize and address the sociocultural context of oppression and unjust distribution of resources and opportunities [[Mallinckrodt et al., 2014](#)]. In statistics textbooks, the use of research vignettes engages the learner around a tangible scenario for identifying independent variables, dependent variables, covariates, and potential mechanisms of change. Many students recall examples in Field's [[2012](#)] popular statistics text: Viagra to teach one-way ANOVA, beer goggles for two-way ANOVA, and bushtucker for repeated measures. What if the research vignettes were more socially responsive?

In this OER, research vignettes will be from recently published articles where:

- the author's identity is from a group where scholarship is historically marginalized (e.g., BIPOC, LGBTQ+, LMIC[low-middle income countries]),

- the research is responsive to issues of justice, equity, inclusion, diversity,
- the lesson's statistic is used in the article, and
- there is sufficient information in the article to simulate the data for the chapter example(s) and practice problem(s); or it is publicly available.

In training for multicultural competence, the saying, “A fish doesn’t know that it’s wet” is often used to convey the notion that we are often unaware of our own cultural characteristics. In recent months and years, there has been an increased awakening to the institutional and systemic racism that our systems are perpetuating. Queuing from the water metaphor, I am hopeful that a text that is recentered in the ways I have described can contribute to *changing the water* in higher education and in the profession of psychology.

## Copyright with Open Access

This book is published under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. This means that this book can be reused, remixed, retained, revised and redistributed (including commercially) as long as appropriate credit is given to the authors. If you remix, or modify the original version of this open textbook, you must redistribute all versions of this open textbook under the same license - CC BY-SA.

A [GitHub open-source repository](#) contains all of the text and source code for the book, including data and images.

# ACKNOWLEDGEMENTS

As a doctoral student at the University of Kansas (1992-2005), I learned that “a foreign language” was a graduation requirement. *Please note that as one who studies the intersections of global, vocational, and sustainable psychology, I regret that I do not have language skills beyond English.* This could have been met with credit from high school my rural, mid-Missouri high school did not offer such classes. This requirement would have typically been met with courses taken during an undergraduate program – but my non-teaching degree in the University of Missouri’s School of Education was exempt from this. The requirement could have also been met with a computer language (fortran, C++) – I did not have any of those either. There was a tiny footnote on my doctoral degree plan that indicated that a 2-credit course, “SPSS for Windows” would substitute for the language requirement. Given that it was taught by my one of my favorite professors, I readily signed up. As it turns out, Samuel B. Green, PhD, was using the course to draft chapters in the textbook [?] that has been so helpful for so many. Unfortunately, Drs. Green (1947 - 2018) and Salkind (2947 - 2017) are no longer with us. I have worn out numerous versions of their text. Another favorite text of mine was Dr. Barbara Byrne’s [2016], “Structural Equation Modeling with AMOS.” I loved the way she worked through each problem and paired it with a published journal article, so that the user could see how the statistical evaluation fit within the larger project/article. I took my tea-stained text with me to a workshop she taught at APA and was proud of the signature she added to it (a little catfur might have fallen out). Dr. Byrne created SEM texts for a number of statistical programs (e.g., LISREL, EQS, MPlus). As I was learning R, I wrote Dr. Byrne, asking if she had an edition teaching SEM/CFA with R. She promptly wrote back, saying that she did not have the bandwidth to learn a new statistics package. We lost Dr. Byrne in December 2020. I am so grateful to these role models for their contributions to my statistical training. I am also grateful for the doctoral students who have taken my courses and are continuing to provide input for how to improve the materials.

The inspiration for training materials that re\*center statistics and research methods came from the [Academics for Black Survival and Wellness Initiative](#). This project, co-founded by Della V. Mosley, Ph.D., and Pearis L. Bellamy, M.S., made clear the necessity and urgency for change in higher education and the profession of psychology.

At very practical levels, I am indebted to SPU’s Library, and more specifically, SPU’s Education, Technology, and Media Department. Assistant Dean for Instructional Design and Emerging Technologies, R. John Robertson, MSc, MCS, has offered unlimited consultation, support, and connection. Senior Instructional Designer in Graphics & Illustrations, Dominic Wilkinson, designed the logo and bookcover. Psychology and Scholarly Communications Librarian, Kristin Hoffman, MLIS, has provided consultation on topics ranging from OERS to citations. I am also indebted to Associate Vice President, Teaching and Learning at Kwantlen Polytechnic University, Rajiv Jhangiani, PhD. Dr. Jhangiani’s text [2019] was the first OER I ever used and I was grateful for

his encouraging conversation.

Financial support for this project has been provided the following:

- *Call to Action on Equity, Inclusion, Diversity, Justice, and Social Responsivity Request for Proposals* grant from the Association of Psychology Postdoctoral and Internship Centers (2021-2022).
- *Diversity Seed Grant*, Office of Inclusive Excellence and Advisory Council for Diversity and Reconciliation (ACDR), Seattle Pacific University.
- *ETM Open Textbook & OER Development Funding*, Office of Education, Technology, & Media, Seattle Pacific University.

# **DATA PREP**



# Chapter 1

## Scrubbing

### [Screencasted Lecture Link](#)

The focus of this chapter is the process of starting with raw data and preparing it for multivariate analysis. To that end, we will address the conceptual considerations and practical steps in “scrubbing and scoring.”

A twist in this lesson is that I am asking you to contribute to the dataset that serves as the basis for the chapter and the practice problems. In the spirit of *open science*, this dataset is available to you and others for your own learning. Before continuing, please take 15-20 minutes to complete the survey titled, [Rate-a-Recent-Course: A ReCentering Psych Stats Exercise](#). The study is approved by the Institutional Review Board at Seattle Pacific University (SPUIRB# 202102011, no expiration). Details about the study, including an informed consent, are included at the link.

### 1.1 Navigating this Lesson

There is about 90 minutes of lecture. If you work through the materials with me it would be good to add another hour.

While the majority of R objects and data you will need are created within the R script that sources the chapter, there are a few that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER’s [introduction](#)

#### 1.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Import data from Qualtrics into R.
- Apply inclusion and exclusion criteria to a dataset.
- Rename variables.
- Create a smaller dataframe with variables appropriate for testing a specific statistical model.
- Use critical data manipulation functions from the *tidyverse* (and *dplyr*) in particular such as *filter()*, *select()*, and *mutate()* to prepare variables.
- Articulate the initial steps in a workflow for scrubbing and scoring data.

### 1.1.2 Planning for Practice

The suggestions for practice will start with this chapter and continue in the next two chapters (Scoring, Data Dx). Using Parent's [2013] AIA (available item analysis) approach to managing missing data, you will scrub-and-score a raw dataset. Options of graded complexity could include:

- Repeating the steps in the chapter with the most recent data from the Rate-A-Recent-Course survey; differences will be in the number of people who have completed the survey since the chapter was written.
- Use the dataset that is the source of the chapter, but score a different set of items that you choose.
- Begin with raw data to which you have access.

### 1.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Parent, M. C. (2013). Handling item-level missing data: Simpler is just as good. *The Counseling Psychologist*, 41(4), 568–600. <https://doi.org/10.1177/0011000012445176>
  - The purpose of Parent's article was to argue that complex and resource-intensive procedures like multiple imputation are unnecessary. Following a simulation that supports his claims, Parent provides some guidelines to follow for the AIA approach.
- Kline, R. B. (2015). Data preparation and psychometrics review. In Principles and Practice of Structural Equation Modeling, Fourth Edition. Guilford Publications. <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4000663>
  - Kline's chapter is my “go-to” for making decisions about preparing data for analysis.

### 1.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
# if(!require(qualtrics)){install.packages('qualtrics')}
# if(!require(tidyverse)){install.packages('tidyverse')}
```

## 1.2 Workflow for Scrubbing

The same workflow guides us through the Scrubbing, Scoring, and Data Dx chapters. In this lesson we focus on downloading data from Qualtrics and determining which cases can be retained for analysis based on inclusion and exclusion criteria.

Here is a narration of the figure:

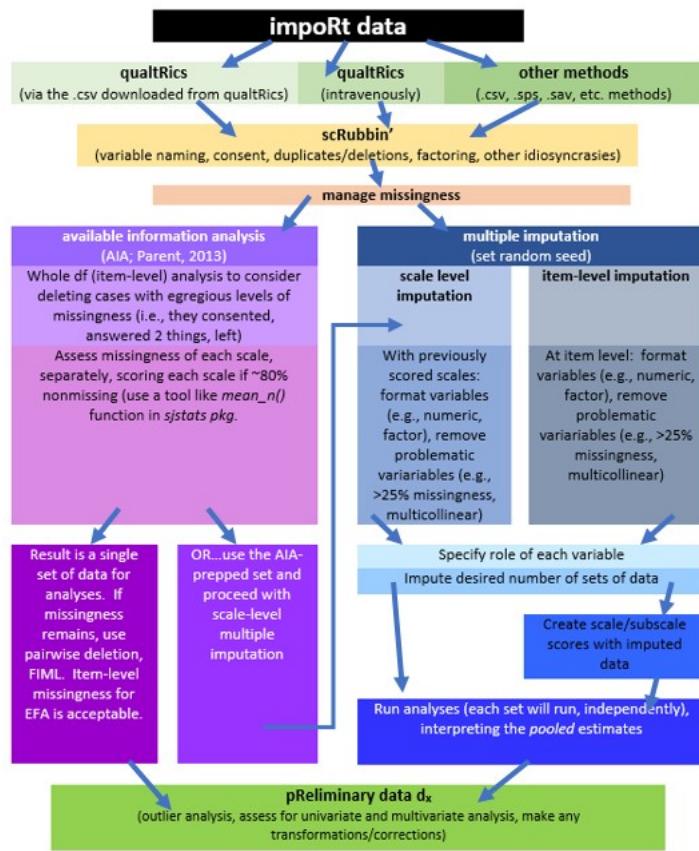


Figure 1.1: An image of a workflow for scrubbing and scoring data.

1. The workflow begins by importing data into R. Most lessons in this series involve simulated data that are created directly in R. Alternatively, data could be:
  - imported “intRavenously” through programs such as Qualtrics,
  - exported from programs such as Qualtrics to another program (e.g., .xlsx, .csv),
  - imported in other forms (e.g., .csv, .sps, .sav).
2. Scrubbing data by
  - variable naming,
  - specifying variable characteristics such as factoring,
  - ensuring that included participants consented to participation,
  - determining and executing the inclusion and exclusion criteria.
3. Conduct preliminary data diagnostics such as
  - outlier analysis
  - assessing for univariate and multivariate analysis
  - making transformations and/or corrections
4. Managing missingness by one of two routes
  - Available information analysis [Parent, 2013] at either the item-level or scale level. The result is a single set of data for analysis. If missingness remains, options include pairwise deletion, listwise deletion, or specifying FIML (when available). Another option is to use multiple imputation.
  - Multiple imputation at either scale level or item-level

## 1.3 Research Vignette

To provide first-hand experience as both the respondent and analyst for the same set of data, you were asked to complete a survey titled, [Rate-a-Recent-Course: A ReCentering Psych Stats Exercise](#). If you haven’t yet completed it, please consider doing so, now. In order to reduce the potential threats to validity by providing background information about the survey, I will wait to describe it until later in the chapter.

The survey is administered in Qualtrics. In the chapter I teach two ways to import Qualtrics data into R. We will then use the data to work through the steps identified in the workflow.

## 1.4 Working the Problem

### 1.4.1 intRavenous Qualtrics

I will demonstrate using a Qualtrics account at my institution, Seattle Pacific University. The only surveys in this account are for the *Recentering Psych Stats* chapters and lessons. The surveys were designed to not capture personally identifying information.

Access credentials for the institutional account, individual user’s account, and survey are essential for getting the survey items and/or results to export into R. The Qualtrics website provides a tutorial for [generating an API token](#).

We need two pieces of information: the **root\_url** and an **API token**. To retrieve these:

- Log into your respective qualtrics.com account
- Select Account Settings
- Choose “Qualtrics IDs” from the user name dropdown

The **root\_url** is the first part of the web address for the Qualtrics account. For our institution it is: *spupsych.az1.qualtrics.com* .

The API token is in the box labeled, “API.” If it is empty, select, “Generate Token.” If you do not have this option, locate the *brand administrator* for your Qualtrics account. They will need to set up your account so that you have API privileges.

*BE CAREFUL WITH THE API TOKEN* This is the key to your Qualtrics accounts. If you leave it in an .rmd file that you forward to someone else or upload to a data repository, this key and the base URL gives access to every survey in your account. If you share it, you could be releasing survey data to others that would violate confidentiality promises in an IRB application.

If you mistakenly give out your API token you can generate a new one within your Qualtrics account and re-protect all its contents.

You do need to change the API key/token if you want to download data from a different Qualtrics account. If your list of surveys generates the wrong set of surveys, restart R, make sure you have the correct API token and try again.

```
# You only need to run this ONCE to draw from the same Qualtrics
# account. If you change Qualtrics accounts you will need to get a
# different token.

# qualtrics::qualtrics_api_credentials(api_key =
# 'mUgPMYSYkiWpMFkwHale1QE5HNmh5LRUaA8d9PDg', base_url =
# 'spupsych.az1.qualtrics.com', overwrite = TRUE, install = TRUE)

# readRenviron('~/Renvironment')
```

*all\_surveys()* generates a dataframe containing information about all the surveys stored on your Qualtrics account.

```
# surveys <- qualtrics::all_surveys()

# View this as an object (found in the right: Environment). Get
# survey id # for the next command If this is showing you the WRONG
# list of surveys, you are pulling from the wrong Qualtrics account
# (i.e., maybe this one instead of your own). Go back and change your
# API token (it saves your old one). Changing the API likely requires
# a restart of R.
```

To retrieve the survey, use the *fetch\_survey()* function.

```
# obtained with the survey ID
#'surveyID' should be the ID from above
#'verbose' prints messages to the R console
#'label', when TRUE, imports data as text responses; if FALSE prints the data as numerical responses
#'convert', when TRUE, attempts to convert certain question types to the 'proper' data type in R
#'force_request', when TRUE, always downloads the survey from the API instead of from a temporary URL
# 'import_id', when TRUE includes the unique Qualtrics-assigned ID;
# since I have provided labels, I want false

# QTRX_df <- qualtrics::fetch_survey(surveyID = 'SV_b2cClqAllGQ6nLU',
# time_zone = NULL, verbose = FALSE, label=FALSE, convert=FALSE,
# force_request = TRUE, import_id = FALSE)

# useLocalTime = TRUE,
```

*It is possible to import Qualtrics data that has been downloaded from Qualtrics as a .csv. I demo this in the Bonus Reel at the end of this lesson.*

In prior versions of this chapter I allowed the chapter to automatically update with “all the new data” each time the OER was re-rendered/built. Because I think this caused confusion, I have decided to save the data in both .csv and .rds versions, then clear my environment, upload the .rds (my personal favorite format) version, and demonstrate the scrubbing techniques with that data. If you continue with data you just downloaded from Qualtrics, you will get different answers than are in the lesson. While I think that continuing with the most current data set is a viable option for a practice problem, it could be confusing. Rather, follow one of the two options below to upload .csv or .rds versions of the data I used in the lesson.

#### 1.4.1.1 Option 1. Upload an .rds file

Because .rds files will retain any formatting information we provide about variables, I like using them. The downside is that you cannot simply open and view them outside of the R environment. Here is the code I used to produce the .rds version of the file. If you want to obtain the same results as I report in the chapter, do NOT run it again.

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(QTRX_df, 'QTRX_df230902.rds')
```

Rather, head to the [MultivModel GitHub](#) site and download the *QTRX\_df230902b.rds* file. Place it in the same folder as the .rmd you are using and run the code below. And actually, I further re-named the file that you will retrieve so that it won’t be over-written.\*

```
QTRX_df <- readRDS("QTRX_df230902b.rds")
```

Occasionally, I have had a student for whom the .rds files don’t seem to work. Uploading a .csv file is an option.

### 1.4.1.2 Option 2. Upload a .csv file

Simply for your information, here is the code I used to produce the .csv version of the file. If you want to obtain the same results as I report in the chapter, do NOT run it again.

```
# write the simulated data as a .csv write.table(QTRX_df,
# file='QTRX_df230902.csv', sep=',', col.names=TRUE, row.names=FALSE)
```

Rather, head to the [MultivModel GitHub](#) site and download the *QTRX\_df230902b.csv* file. Place it in the same folder as the .rmd you are using and run the code below. *And actually, I further re-named the file that you will retrieve so that it won't be over-written.*

```
# bring back the simulated dat from a .csv file QTRX_df <-
# read.csv('QTRX_df230902b.csv', header = TRUE)
```

You need not do both. That is, either download-and-import either the .rds or .csv file.

### 1.4.2 About the *Rate-a-Recent-Course Survey*

As a teaching activity for the ReCentering Psych Stats OER, the topic of the survey was selected to be consistent with the overall theme of OER. Specifically, the purpose of this study is to understand the campus climate for students whose identities make them vulnerable to bias and discrimination. These include students who are Black, non-Black students of color, LGBTQ+ students, international students, and students with disabilities.

Although the dataset should provide the opportunity to test a number of statistical models, one working hypothesis that framed the study is that there will be a greater sense of belonging and less bias and discrimination when there is similar representation (of identities that are often marginalized) in the instructional faculty and student body. Termed, “structural diversity” [Lewis and Shah, 2019] this is likely an oversimplification. In fact, an increase in diverse representation without attention to interacting factors can increase hostility on campus [Hurtado, 2007]. Thus, we included the task of rating of a single course relates to the larger campus along the dimensions of belonging and bias/discrimination. For example, if a single class has higher ratings on issues of inclusivity, diversity, and respect, we would expect that sentiment to be echoed in the broader institution.

Our design has notable limitations. You will likely notice that we ask about demographic characteristics of the instructional staff and classmates in the course rated, but we do not ask about the demographic characteristics of the respondent. In making this decision, we likely lose important information; Iacovino and James [2016] have noted that White students perceive campus more favorably than Black student counterparts. We made this decision to protect the identity of the respondent. As you will see when we download the data, if a faculty member asked an entire class to take the survey, the datestamp and a handful of demographic identifiers could very likely identify a student. In certain circumstances, this might be risky in that private information (i.e., gender nonconformity, disclosure of a disability) or course evaluation data could be related back to the student.

Further, the items that ask respondents to *guess* the identities of the instructional staff and classmates are limited, and contrary to best practices in survey construction that recommend providing

the option of a “write-in” a response. After consulting with a diverse group of stakeholders and subject matter experts (and revising the response options numerous times) I have attempted to center anti-Black racism in the U.S. [Mosley et al., 2021, 2020, Singh, 2020]. In fact, the display logic does not present the race items when the course is offered outside the U.S. There are only five options for race: *biracial/multiracial*, *Black*, *non-Black person(s) of color*, *White*, and *I did not notice* (intended to capture a color-blind response). One unintended negative consequence of this design is that the response options could contribute to *colorism* [Adames et al., 2021, Capielo Rosario et al., 2019]. Another possibility is that the limited options may erase, or make invisible, other identities. At the time that I am writing the first draft of this chapter, the murder of six Asian American women in Atlanta has just occurred. The Center for the Study of Hate and Extremeism has documented that while overall hate crimes dropped by 7% in 2020, anti-Asian hate crimes reported to the police in America’s largest cities increased by 149% [noa, a]. These incidents have occurred not only in cities, but in our neighborhoods and on our campus [Kim, 2021b,a, noa, b]. While this survey is intended to assess campus climate as a function of race, it unfortunately does not distinguish between many identities that experience marginalization.

In parallel, the items asking respondents to identify characteristics of the instructional staff along dimensions of gender, international status, and disability are “large buckets” and do not include “write-in” options. Similarly, there was no intent to cause harm by erasing or making invisible individuals whose identities are better defined by different descriptors. Further, no write-in items were allowed. This was also intentional to prevent potential harm caused by people who could leave inappropriate or harmful comments.

### 1.4.3 The Codebook

In order to scrub-and-score a survey, it is critical to know about its content, scoring directions for scales/subscales, and its design. A more complete description of the survey design elements is (or will be) available in the *Recentering Psych Stats: Psychometric OER*. The review in this chapter provides just-enough information to allow us to make decisions about which items to retain and how to score them. When they are well-written, information in the **IRB application** and **pre-registration** can be helpful in the scrubbing and scoring process.

Let’s look “live” at the survey. In Qualtrics it is possible to *print* a PDF that looks very similar to its presentation when someone is taking it. You can access that static version [here](#).

We can export a **codebook**, that is, a Word (or PDF) version of the survey with all the coding. In Qualtrics the protocol is: Survey/Tools/ImportExport/Export Survey to Word. Then select all the options you want (especially “Show Coded Values”). A tutorial provided by Qualtrics can be found [here](#). This same process can be used to print the PDF example I used above.

It is almost impossible to give this lecture without some reference to Qualtrics and the features used in Qualtrics. An import of raw data from Qualtrics into R can be nightmare in that the Qualtrics-assigned variable names are numbers (e.g., QID1, QID2) – but often out of order because the number is assigned when the question is first created. If the survey is reordered, the numbers get out of sequence.

Similarly, values for Likert-type scales can also get out of order if the scale anchors are revised (which is common to do).

I recommend providing custom variable names and recode values directly in Qualtrics before exporting them into R. A Qualtrics tutorial for this is provided [here](#). In general, consider these

qualities when creating variable names:

- Brevity: historically, SPSS variable names could be a maximum of 8 characters.
- Intuitive: although variables can be renamed in R (e.g., for use in charts and tables), it is helpful when the name imported from Qualtrics provides some indication of what the variable is.
- Systematic: start items in a scale with the same stem, followed by the item number – ITEM1, ITEM2, ITEM3.

The Rate-a-Recent-Course survey was written using some special features in Qualtrics. These include

- Display logic
  - Items that are U.S.-centric are only shown if the respondent is taking a course from an institution in the U.S. is a student in the U.S.
- Loop and merge
  - Because course may have multiple instructional staff, the information asking about demographic characteristics of the instructors is repeated according to the number input by the respondent
- Random presentation of the 30 items asking about campus climate for the five groups of students
  - Although this might increase the cognitive load of the survey, this helps “spread out” missingness for respondents who might tire of the survey and stop early
- Rank ordering of the institutional level (department, school/faculty, campus/university) to which the respondent feels most connected

Looking at the QTRX\_df, *StartDate* thru *UserLanguage* are metadata created by Qualtrics. The remaining variables and associated value labels are in the [codebook](#).

## 1.5 Scrubbing

With a look at our survey, codebook, and imported data, we now get to the business of scrubbing (deleting those who did not give consent, deleting previews, etc.). This level of “scrubbing” precedes the more formal detection of outliers.

### 1.5.1 Tools for Data Manipulation

The next stages will provide some experience manipulating data with **dplyr** from the **tidyverse**.

The **tidyverse** is a system of packages (i.e., when you download the tidyverse, you download all its packages/members) for data manipulation, exploration and visualization. The packages in the tidyverse share a common design philosophy. These were mostly developed by Hadley Wickham,

but more recently, more designers are contributing to them. Tidyverse packages are intended to make statisticians and data scientists more productive by guiding them through workflows that facilitate communication and result in reproducible work products. Fundamentally, the tidyverse is about the connections between the tools that make the workflow possible. Critical packages in the tidyverse include:

- **dplyr**: data manipulation: mutate, select, filter, summarize, arrange
- **ggplot2**: extravagant graphing
- **tibble**: a *tibble* is a dataframe that provides the user with more (and less) control over the data.
- **readr**: gives access to “rectangular data” like .csv and tables
- **tidyverse**: tidy data is where each variable is a column, each observation is a row, each value is a cell (duh). **tidyverse**’s contributions are gather(wide to long) and spread(long to wide) as well as separate, extract, unite.
- **purrr**: facilitates working with functions and vectors. For example, if you write a function, using purrr may help you replace loops with code that is more efficient and intuitive.

The tidyverse is ever-evolving – so check frequently for updates and troubleshooting.

A handy cheatsheet for data transformation is found [here](#).

### 1.5.2 Inclusion and Exclusion Criteria

For me, the first pass at scrubbing is to eliminate the obvious. In our case this includes *previews* and respondents who did not consent to continue. Previews are the researcher-initiated responses usually designed to proofread or troubleshoot survey problems. There could be other first-pass-deletions, such as selecting response between certain dates.

I think these first-pass deletions, especially the ones around consent, are important to do as soon as possible. Otherwise, we might delete some of the variables (e.g., timestamps, consent documentation, preview status) and neglect to delete these cases later in the process.

We are here in the workflow:

We can either update the existing df (by using the same object), or creating a new df from the old. Either works. In my early years, I tended to create lots of new objects. As I have gained confidence in myself and in R, I’m inclined to update the existing df. Why? Because unless you write the object as an outfile (using the same name for the object as for the filename – which I do not recommend), the object used in R does not change the source of the dat. Therefore, it is easy to correct early code and it keeps the global environment less cluttered.

In this particular survey, the majority of respondents will take the survey because they clicked an *anonymous* link provided by Qualtrics. Another Qualtrics distribution method is e-mail. At the time of this writing, we have not recruited by e-mail, but it is possible we could do so in the future. What we should not include, though, are *previews*. These are the times when the researcher is self-piloting the survey to look for errors and to troubleshoot.

```
# the filter command is used when we are making inclusion/exclusion
# decisions about rows != means do not include cases with 'preview'
```

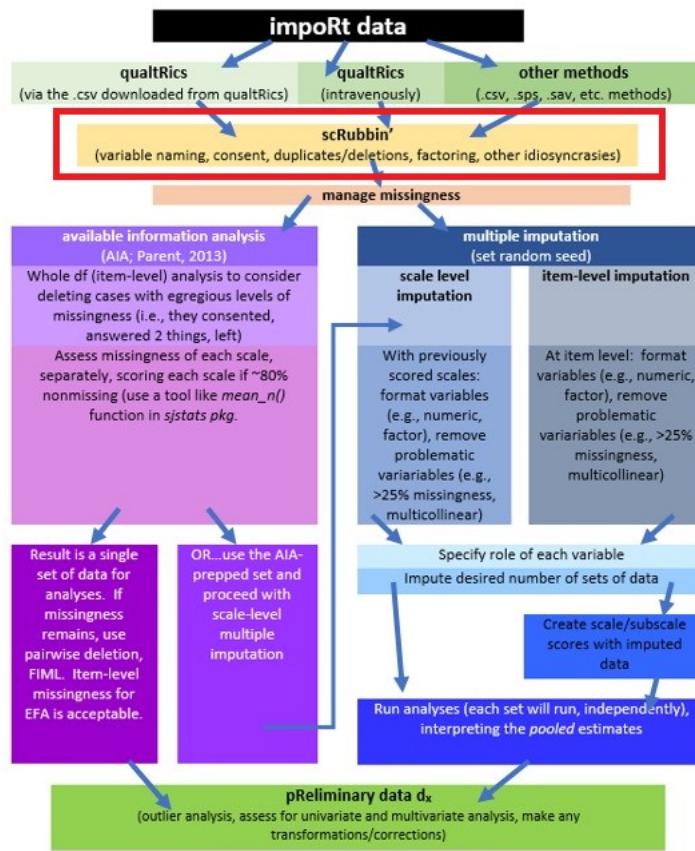


Figure 1.2: An image of a workflow for scrubbing and scoring data.

```

QTRX_df <- dplyr::filter(QTRX_df, DistributionChannel != "preview")

# FYI, another way that doesn't use tidyverse, but gets the same
# result QTRX_df <- QTRX_df[!QTRX_df$DistributionChannel ==
# 'preview',]

```

APA Style, and in particular the Journal Article Reporting Standards (JARS) for quantitative research specify that we should report the frequency or percentages of missing data. We would start our counting *after* eliminating the previews.

```

# I created an object that lists how many rows/cases remain. I used
# inline text below to update the text with the new number
nrow(QTRX_df)

```

[1] 107

#### CAPTURING RESULTS FOR WRITING IT UP:

Data screening suggested that 107 individuals opened the survey link.

Next let's filter in only those who consented to take the survey. Because Qualtrics discontinued the survey for everyone who did not consent, we do not have to worry that their data is unintentionally included, but it can be useful to mention the number of non-consenters in the summary of missing data.

```

# == are used
QTRX_df <- dplyr::filter(QTRX_df, Consent == 1)
nrow(QTRX_df)

```

[1] 83

#### CAPTURING RESULTS FOR WRITING IT UP:

Data screening suggested that 107 individuals opened the survey link. Of those, 83 granted consent and proceeded into the survey items.

In this particular study, the categories used to collect information about race/ethnicity were U.S.-centric. Thus, they were only shown if the respondent indicated that the course being rated was taught by an institution in the U.S. Therefore, an additional inclusion criteria for this specific research model should be that the course was taught in the U.S.

```

QTRX_df <- dplyr::filter(QTRX_df, USinst == 0)
nrow(QTRX_df)

```

```
[1] 69
```

#### CAPTURING RESULTS FOR WRITING IT UP:

Data screening suggested that 107 individuals opened the survey link. Of those, 83 granted consent and proceeded into the survey items. A further inclusion criteria was that the course was taught in the U.S; 69 met this criteria.

#### 1.5.3 Renaming Variables

Even though we renamed the variables in Qualtrics, the loop-and-merge variables were auto-renamed such that they each started with a number. I cannot see how to rename these from inside Qualtrics. A potential problem is that, in R, when variable names start with numbers, they need to be surrounded with single quotation marks. I find it easier to rename them now. I used “i” to start the variable name to represent “instructor.”

The form of the *rename()* function is this: df\_named <- rename(df\_raw, NewName1 = OldName1)

```
QTRX_df <- dplyr::rename(QTRX_df, iRace1 = "1_iRace", iRace2 = "2_iRace",
                           iRace3 = "3_iRace", iRace4 = "4_iRace", iRace5 = "5_iRace", iRace6 = "6_iRace",
                           iRace7 = "7_iRace", iRace8 = "8_iRace", iRace9 = "9_iRace", iRace10 = "10_iRace")
```

Also in Qualtrics, it was not possible to rename the variable (formatted with sliders) that asked respondents to estimate the proportion of classmates in each race-based category. Using the code-book, we can do this now. I will use “cm” to precede each variable name to represent “classmates.”

```
QTRX_df <- dplyr::rename(QTRX_df, cmBiMulti = Race_10, cmBlack = Race_1,
                           cmNBPoC = Race_7, cmWhite = Race_8, cmUnsure = Race_2)
```

Let’s also create an ID variable (different from the lengthy Qualtrics-issued ID) and then move it to the front of the distribution.

```
# Opening the tidyverse so that I can use pipes
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr    1.3.0
v purrr     1.0.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```

QTRX_df <- QTRX_df %>%
  dplyr::mutate(ID = row_number())

# moving the ID number to the first column; requires
QTRX_df <- QTRX_df %>%
  dplyr::select(ID, everything())

```

#### 1.5.4 Downsizing the Dataframe

Although researchers may differ in their approach, my tendency is to downsize the df to the variables I will be using in my study. These could include variables in the model, demographic variables, and potentially auxiliary variables (i.e., variables not in the model, but that might be used in the case of multiple imputation).

This particular survey did not collect demographic information, so that will not be used. The model that I will demonstrate in this research vignette examines the the respondent's perceived campus climate for students who are Black, predicted by the the respondent's own campus belonging, and also the *structural diversity* [Lewis and Shah, 2019] proportions of Black students in the classroom and BIPOC (Black, Indigenous, and people of color) instructional staff.

*I would like to assess the model by having the instructional staff variable to be the %Black instructional staff. At the time that this lecture is being prepared, there is not sufficient Black representation in the staff to model this.*

The `select()` function can let us list the variables we want to retain.

```

# You can use the ':' to include all variables from the first to last
# variable in any sequence; I could have written this more
# efficiently. I just like to 'see' my scales and clusters of
# variables.

Model_df <- (dplyr::select(QTRX_df, ID, iRace1, iRace2, iRace3, iRace4,
  iRace5, iRace6, iRace7, iRace8, iRace9, iRace10, cmBiMulti, cmBlack,
  cmNBPoC, cmWhite, cmUnsure, Belong_1:Belong_3, Blst_1:Blst_6))

```

It can be helpful to save outfile of progress as we go along. Here I save this raw file. I will demonstrate how to save both .rds and .csv files.

```

# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Model_df, 'BlackStntsModel230902.rds') code to import
# that model we just saved Model_df <-
# readRDS('BlackStntsModel230902.rds')

# write the simulated data as a .csv write.table(Model_df,
# file='BlackStntsModel230902.csv', sep=',', col.names=TRUE,
# row.names=FALSE) bring back the simulated data from a .csv file
# Model_df <- read.csv('BlackStntsModel230902.csv', header = TRUE)

```

## 1.6 Toward the APA Style Write-up

Because we have been capturing the results as we have worked the problem, our results section is easy to assemble.

### 1.6.1 Method/Procedure

Data screening suggested that 107 individuals opened the survey link. Of those, 83 granted consent and proceeded into the survey items. A further inclusion criteria was that the course was taught in the U.S; 69 met this criteria.

## 1.7 Practice Problems

Starting with this chapter, the practice problems for this and the next two chapters (i.e., Scoring, Data Dx) are intended to be completed in a sequence. Whatever practice option(s) you choose, please

- Use raw data that has some missingness (as a last resort you could manually delete some cells),
- Includes at least 3 independent/predictor variables
  - these could be categorically or continuously scaled
  - at least one variable should require scoring.
- Include at least 1 dependent variable
  - at this point in your learning it should be continuously scaled

The three problems below are listed in the order of graded complexity. If you are just getting started, you may wish to start with the first problem. If you are more confident, choose the second or third option. You will likely encounter challenges that were not covered in this chapter. Search for and try out solutions, knowing that there are multiple paths through the analysis.

### 1.7.1 Problem #1: Rework the Chapter Problem

Because the *Rate-a-Recent-Course* survey remains open, it is quite likely that there will be more participants who have taken the survey since this chapter was last updated. If not – please encourage a peer to take it. Even one additional response will change the results. This practice problem encourages you to rework the chapter, as written, with the updated data from the survey.

### 1.7.2 Problem #2: Use the *Rate-a-Recent-Course* Survey, Choosing Different Variables

Before starting this option, choose a minimum of three variables from the *Rate-a-Recent-Course* survey to include in a simple statistical model. Work through the chapter making decisions that

are consistent with the research model you have proposed. There will likely be differences at several points in the process. For example, you may wish to include (not exclude) data where the rated-course was offered by an institution outside the U.S. Different decisions may involve an internet search for the R script you will need as you decide on inclusion and exclusion criteria.

### 1.7.3 Problem #3: Other data

Using raw data for which you have access, use the chapter as a rough guide. Your data will likely have unique characteristics that may involved searching for solutions beyond this chapter/OER.

### 1.7.4 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Assignment Component	Points Possible	Points Earned
1. Specify a research model that includes three predictor variables (continuously or categorically scaled) and one dependent (continuously scaled) variable	5	_____
2. Import data	5	_____
3. Include only those who consented*	5	_____
4. Apply exclusionary criteria *	5	_____
5. Rename variables to be sensible and systematic *	5	_____
6. Downsize the dataframe to the variables of interest	5	_____
7. Provide an APA style write-up of these preliminary steps	5	_____
8. Explanation to grader	5	_____
<b>Totals</b>	<b>40</b>	_____

\* If your dataset does not require these steps, please provide example code that uses variables in your dataset. For example, for the inclusion or exclusion criteria, provide an example of how to filter in (or out) any variable on the basis of one of the response options. Once demonstrated, hashtag it out and rerun your script with those commands excluded.

A *homeworked example* for the Scrubbing, Scoring, and DataDx lessons (combined) follows the [Data Dx](#) lesson.

## 1.8 Bonus Track:

### 1.8.1 Importing data from an exported Qualtrics .csv file

The lecture focused on the “intRavenous” import. It is also possible to download the Qualtrics data in a variety of formats (e.g., CSV, Excel, SPSS). Since I got started using files with the CSV extension (think “Excel” lite), that is my preference.

In Qualtrics, these are the steps to download the data: Projects/YOURsurvey/Data & Analysis/Export & Import/Export data/CSV/Use numeric values

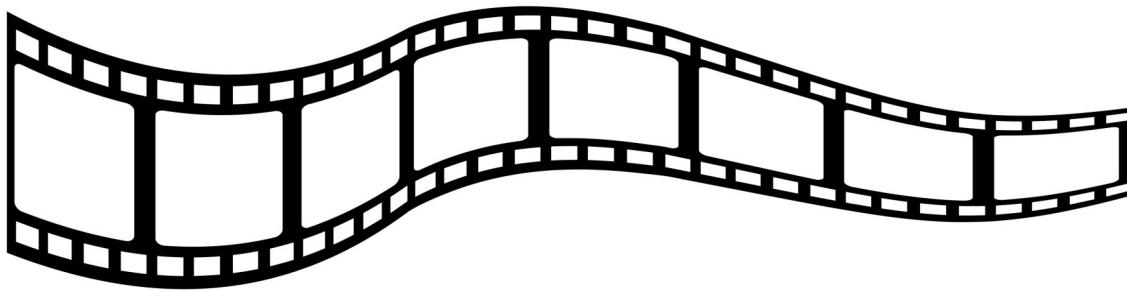


Figure 1.3: Image of a filmstrip

I think that it is critical that to save this file in the same folder as the .rmd file that you will use with the data.

R is sensitive to characters used in filenames. As downloaded, my Qualtrics .csv file had a long name with spaces and symbols that are not allowed. Therefore, I gave it a simple, sensible, filename, “ReC\_Download210319.csv”. An idiosyncracy of mine is to datestamp filenames. I use two-digit representations of the year, month, and date so that if the letters preceding the date are the same, the files would alphabetize automatically.

```
library(qualtRics)
QTRX_csv <- read_survey("ReC_Download210319.csv", strip_html = TRUE, import_id = FALSE,
  time_zone = NULL, legacy = FALSE)
```

```
-- Column specification -----
cols(
  .default = col_double(),
  StartDate = col_datetime(format = ""),
  EndDate = col_datetime(format = ""),
  RecordedDate = col_datetime(format = ""),
  ResponseId = col_character(),
  DistributionChannel = col_character(),
  UserLanguage = col_character(),
  Virtual = col_number(),
  `^5_iPronouns` = col_logical(),
  `^5_iGenderConf` = col_logical(),
  `^5_iRace` = col_logical(),
  `^5_iUS` = col_logical(),
  `^5_iDis` = col_logical(),
  `^6_iPronouns` = col_logical(),
  `^6_iGenderConf` = col_logical(),
  `^6_iRace` = col_logical(),
  `^6_iUS` = col_logical(),
```

```
`6_iDis` = col_logical(),
`7_iPronouns` = col_logical(),
`7_iGenderConf` = col_logical(),
`7_iRace` = col_logical()
# ... with 17 more columns
)
i Use `spec()` for the full column specifications.
```

Although minor tweaking may be required, the same script above should be applicable to this version of the data.

# Chapter 2

## Scoring

### [Screencasted Lecture Link](#)

The focus of this chapter is to continue the process of scrubbing-and-scoring. We continue with the raw data we downloaded and prepared in the prior chapter. In this chapter we analyze and manage missingness, score scales/subscales, and represent our work with an APA-style write-up. To that end, we will address the conceptual considerations and practical steps in this process.

### 2.1 Navigating this Lesson

There is about 1 hour and 20 minutes of lecture. If you work through the materials with me it would be good to add another hour.

While the majority of R objects and data you will need are created within the R script that sources the chapter, there are a few that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

#### 2.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Recognize the key components of data loss mechanisms (MCAR, MAR, MNAR), including how to diagnose MCAR.
- Interpret missingness figures produced by packages such as *mice*.
- Articulate a workflow for scrubbing and scoring data.
- Use critical data manipulation functions from *dplyr* including *filter()*, *select()*, and *mutate()* to prepare variables.
- Interpret code related to missingness (i.e., “is.na”, “!is.na”) and the pipe (%>%)

#### 2.1.2 Planning for Practice

The suggestions for practice continue from the prior chapter. The assignment in the prior chapter involved downloading a dataset from Qualtrics and the “scrubbing” it on the basis of inclusion

and exclusion criteria. Using that same data, the practice suggestions in this chapter will continue to use Parent's [2013] AIA approach to managing missing data, to score the variables of interest. Options of graded complexity could include:

- Repeating the steps in the chapter with the most recent data from the Rate-A-Recent-Course survey; differences will be in the number of people who have completed the survey since the chapter was written.
- Use the dataset that is the source of the chapter, but score a different set of items that you choose.
- Begin with raw data to which you have access.

### 2.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Enders, C. K. (2010). Applied missing data analysis (2010-13190-000). Guilford Press.
  - Enders' text continues to be the comprehensive “go-to” source for examining and managing missing data.
- Kline, R. B. (2015). Data preparation and psychometrics review. In Principles and Practice of Structural Equation Modeling, Fourth Edition. Guilford Publications. <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4000663>
  - Kline’s chapter is my “go-to” for making decisions about preparing data for analysis.
- Parent, M. C. (2013). Handling item-level missing data: Simpler is just as good. *The Counseling Psychologist*, 41(4), 568–600. <https://doi.org/10.1177/0011100012445176>
  - The purpose of Parent’s article was to argue that complex and resource-intensive procedures like multiple imputation are unnecessary. Following a simulation that supports his claims, Parent provides some guidelines to follow for the AIA approach.

### 2.1.4 Packages

The packages used in this lesson are embedded in this code. When the hashtags are removed, the script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# if(!require(tidyverse)){install.packages('tidyverse')}
# if(!require(psych)){install.packages('psych')}
# if(!require(mice)){install.packages('mice')}
# if(!require(sjstats)){install.packages('sjstats')}
# if(!require(formattable)){install.packages('formattable')}
```

## 2.2 Workflow for Scoring

The following is a proposed workflow for preparing data for analysis.

The same workflow guides us through the Scrubbing, Scoring, and Data Dx chapters. At this stage in the chapter we are still scrubbing as we work through the item-level and whole-level portions of the AIA (left side) of the chart.

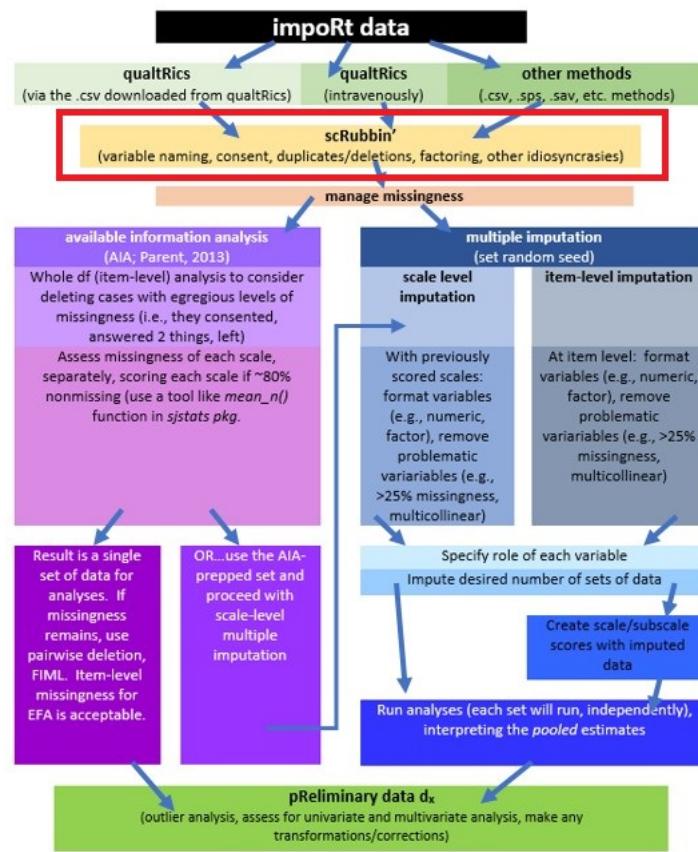


Figure 2.1: An image of our stage in the workflow for scrubbing and scoring data.

## 2.3 Research Vignette

The research vignette comes from the survey titled, [Rate-a-Recent-Course: A ReCentering Psych Stats Exercise](#) and is explained in the prior chapter. In the prior chapter we conducted super-preliminary scrubbing of variables that will allow us to examine the respondent's perceived campus climate for students who are Black, predicted by the the respondent's own campus belonging, and also the *structural diversity* proportions of Black students in the classroom and the BIPOC instructional staff. At present, I see this as a parallel mediation. That is, the perceived campus climate for Black students will be predicted by the respondent's sense of belonging, through the proportion of Black classmates and BIPOC (Black, Indigenous, and people of color)instructional staff.

I would like to assess the model by having the instructional staff variable to be the percent of Black instructional staff. At the time that this lecture is being prepared, there is insufficient representation of Black faculty to model this.

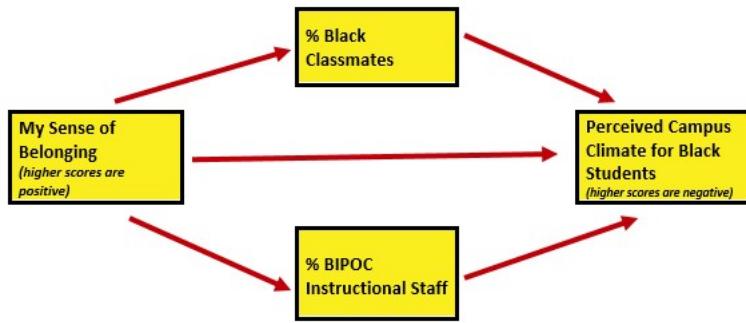


Figure 2.2: An image of the statistical model for which we are preparing data.

First, though, let's take a more conceptual look at issues regarding missing data. We'll come back to details of the survey as we work with it.

## 2.4 On Missing Data

On the topic of missing data, we follow the traditions in most textbooks. We start by considering *data loss mechanisms* and options for *managing missingness*.

Although the workflow I recommend is fairly straightforward, the topic is not. Quantitative psychologists have produced volumes of research that supports and refutes all of these issues in detail. An in-depth review of this is found in Enders' [2010] text.

### 2.4.1 Data Loss Mechanisms

We generally classify missingness in data in three different ways [Kline, 2016b, Parent, 2013]:

**Missing completely at random (MCAR)** is the ideal case (and often unrealistic in actual data). For variable  $Y$  this means that

- Missingness is due to a factor(s) completely unrelated to the missing data. Stated another way:
  - Missing observations differ from the observed scores only by chance; that is, whether scores on  $Y$  are missing or not missing is unrelated to  $Y$  itself
- The presence versus absence of data on  $Y$  is unrelated to all other variables in the dataset. That is, the nonmissing data are just a random sample of scores that the researcher would have analyzed had the data been complete. We might think of it as *haphazard* missing.
  - A respondent is interrupted, looks up, looks down, and skips an item.
  - A computer glitch causes spotty missingness – unrelated to any particular variable.

MCAR is the ideal state because results from it should not be biased as a function of the missingness.

**Missing at random (MAR)** missing data arise from a process that is both measured and predictable in a particular sample. *Admittedly the use of “random” in this term is odd, because, by definition, the missingness is not random.*

Restated:

1. Missingness on Y is unrelated to Y itself, but
2. Missingness is on Y is correlated with other variables in the data set.

Example: Men are less likely to respond to questions about mental health than women, but among men, the probability of responding is unrelated to their true mental health status.

Kline [2016b] indicated that information loss due to MAR is potentially recoverable through imputation where missing scores are replaced by predicted scores. The predicted scores are generated from other variables in the data set that predict missingness on Y. If the strength of that prediction is reasonably strong, then results on Y after imputation may be relatively unbiased. In this sense, the MAR pattern is described as *ignorable* with regard to potential bias. Two types of variables can be used to predict the missing data

1. variables that are in the prediction equation, and
2. *auxiliary* variables (i.e., variables in the dataset that are not in the prediction equation).

Parent [2013] noted that multiple imputation and expectation maximization have frequently been used to manage missingness in MAR circumstances.

**Missing not at random (MNAR)** is when the presence versus absence of scores on Y depend on Y itself. This is *non-ignorable*.

For example, if a patient drops out of a medical RCT because there are unpleasant side effects from the treatment, this discomfort is not measured, but the data is missing due to a process that is unknown in a particular data set. Results based on *complete cases only* can be severely biased when the data loss pattern is MNAR. That is, a treatment may look more beneficial than it really is if data from patients who were unable to tolerate the treatment are lost.

Parent [2013] described MNAR a little differently – but emphasized that the systematic missingness would be related to a variable outside the dataset. Parent provided the example of items written in a manner that may be inappropriate for some participants (e.g., asking women about a relationship with their boyfriend/husband, when the woman might be in same gender relationship). If there were not demographic items that could identify the bias, this would be MNAR. Parent strongly advises researchers to carefully proofread and pilot surveys to avoid MNAR circumstances.

Kline [2016b] noted that the choice of the method to deal with the incomplete records can make a difference in the results, and should be made carefully.

#### 2.4.2 Diagnosing Missing Data Mechanisms

The bad news is that we never really know (with certainty) the type of missing data mechanism in our data. The following tools can help understand the mechanisms that contribute to missingness.

- Missing data analyses often includes correlations that could predict missingness.
- Little and Rubin [2002] proposed a multivariate statistical test of the MCAR assumption that simultaneously compares complete versus incomplete cases on  $Y$  across all other variables. If this comparison is significant, then the MCAR hypothesis is rejected.
  - To restate: we want a non-significant result; and we use the sometimes-backwards-sounding NHST (null hypothesis significance testing) language, “MCAR cannot be rejected.”
- MCAR can also be examined through a series of  $t$  tests of the cases that have missing scores on  $Y$  with cases that have complete records on other variables. Unfortunately, sample sizes contribute to problems with interpretation. With low samples, they are underpowered; in large samples they can flag trivial differences.

If MCAR is rejected, we are never sure whether the data loss mechanism is MAR or MNAR. There is no magical statistical “fix.” Kline [2016b] wrote, “About the best that can be done is to understand the nature of the underlying data loss pattern and accordingly modify your interpretation of the results” (p. 85).

### 2.4.3 Managing Missing Data

There are a number of approaches to managing missing data. Here is a summary of the ones most commonly used.

- **Listwise deletion** (aka, Complete Case Analysis) If there is a missing score on any variable, that case is excluded from **all** analyses.
- **Pairwise deletion** Cases are excluded only if they have missing data on variables involved in a particular analysis. AIA is a variant of pair-wise deletion, but it preserves as much data as possible with person-mean imputation at the scale level.
- **Mean/median substitution** Mean/median substitution replaces missing values with the mean/median of that particular variable. While this preserves the mean of the dataset, it can cause bias by decreasing variance. For example, if you have a column that has substantial of missingness and you replace each value with the same, fixed, mean, the variability of that variable has just been reduced. A variation on this is a **group-mean substitution** where the missing score in a particular group (e.g., women) is replaced by the group mean.
- **Full information maximum likelihood (FIML)** A *model-based method* that takes the researcher’s model as the starting point. The procedure partitions the cases in a raw data file into subsets, each with the same pattern of missing observations, including none (complete cases). Statistical information (e.g., means, variances) is extracted from each subset so all case are retained in the analysis. Parameters for the researcher’s model are estimated after combining all available information over the subsets of cases.
- **Multiple imputation** A *data based method* that works with the whole raw data file (not just with the observed variables that comprise the researcher’s model). Multiple imputation assumes that data are MAR (remember, MCAR is the more prestigious one). This means that researchers assume that missing values can be replaced by predictions derived from the observable portion fo the dataset.

- Multiple datasets (often 5 to 20) are created where missing values are replaced via a randomized process (so the same missing value [item 4 for person A] will likely have different values for each dataset).
- The desired analysis(es) is conducted simultaneously/separately for each of the imputed sets (so if you imputed 5 sets and wanted a linear regression, you get 5 linear regressions).
- A *pooled analysis* uses the point estimates and the standard errors to provide a single result that represents the analysis.

#### 2.4.4 Available Information Analysis (AIA)

Parent [2013] has created a set of recommendations that help us create a streamlined workflow for managing missing data. After evaluating three approaches to managing missingness (AIA, mean substitution, and multiple imputation) Parent concluded that in datasets with (a) low levels of missingness, (b) a reasonable sample size, and (c) adequate internal reliability of measures, these approaches had similar results.

Further, in simulation studies where there was (a) low sample size ( $n = 50$ ), (b) weak associations among items, and (c) a small number of missing items, AIA was equivalent to multiple imputation. Even in cases where the data conditions were the “best” (i.e.,  $N = 200$ , moderate correlations, at least 10 items), even 10% missingness (overall) did not produce notable difference among the methods. That is, means, standard errors, and alphas were similar across the methods (AIA, mean substitution, multiple imputation).

AIA is an older method of handling missing data that, as its name suggests, uses the *available data* for analysis and excludes missing data points only for analyses in which the missing data point would be directly involved. This means

- In the case of research that uses multiple item scales, and analysis takes place at the scale level
  - AIA is used to generate **mean** scores for the scale using the available data without substituting or imputing values;
  - This method generally produces a fairly complete set of scale-level data where
    - \* pairwise deletion (the whole row/case/person is skipped) can be used where there will be multiple analyses using statistics (e.g., correlations, t-tests, ANOVA) where missingness is not permitted
    - \* FIML can be specified in path analysis and CFA/SEM (where item-level data is required), and
    - \* some statistics, such as principal components analysis and principal axis factoring (item-level analyses) permit missing data,
  - Of course, the researcher could still impute data, but why...

Parent's [2013] recommendations:

- Scale scores should be first calculated as a *mean* (average) not a sum. Why?
  - Calculating a “sum” from available data will result in automatically lower scores in cases where there is missingness.

- If a sum is required (i.e., because you want to interpret some clinical level of something), calculate the mean first, do the analyses, then transform the results back into the whole-scale equivalent (multiply the mean by the number of items) for any interpretation.
- For R script, do not write the script ( $[item1 + item2 + item3]/3$ ) because this will return an empty entry for participants missing data (same problem as if you were to use sum). There are several functions for properly computing a mean; I will demo the *mean\_n()* function from *sjstats* package because it allows us to simultaneously specify the tolerance level (next item).
- Determine your *tolerance* for missingness (20% seems to be common, although you could also look for guidance in the test manual/article). Then
  - Run a “percent missingness” check on the level of analysis (i.e., total score, scale, or subscale) you are using. If you are using a total scale score, then check to see what percent is missing across all the items in the whole scale. In contrast, if you are looking at subscales, run the percent missing at that level.
  - Parent [2013] advised that the tolerance levels should be made mindfully. A four-item scale with one item missing, won’t meet the 80% threshold, so it may make sense to set a 75% threshold for this scale.
- “Clearly and concisely detail the level of missingness” in papers [Parent, 2013, p. 595]. This includes
  - tolerance level for missing data by scale or subscale (e.g., 80% or 75%)
  - the number of missing values out of all data points on that scale for all participants and the maximum by participant (e.g., “For Scale X, a total of # missing data points out of ### were observed with no participant missing more than a single point.”)
  - verify a manual inspection of missing data for obvious patterns (e.g., abnormally high missing rates for only one or two items). This can be accomplished by requesting frequency output for the items and checking the nonmissing data points for each scale, ensuring there are no abnormal spikes in missingness (looking for MNAR).
- Curiously, Parent [2013] does not recommend that we run all the diagnostic tests. However, because recent reviewers have required them of me, I will demonstrate a series of them.
- Reducing missingness starts at the survey design – make sure that all people can answer all items (i.e., relationship-related items may contain heterosexist assumptions...which would result in an MNAR circumstance)

Very practically speaking, Parent’s [2013] recommendations follow us through the entire data analysis process.

## 2.5 Working the Problem

### 2.5.1 Variable Planning and Preparation

In the **Scrubbing lesson** we imported the data from Qualtrics and applied the broadest levels of inclusion (e.g., the course rated was offered from an institution in the U.S., the respondent consented to participation) and exclusion (e.g., the survey was not a preview). We then downsized the survey

to include the variables we will use in our statistical model. We then saved the data in .csv and .rds file.

Presuming that you are working along with me in an .rmd file and have placed that file in the same folder as this .rmd file, the following code should read the data into your environment.

I use *different* names for the object/df in my R environment than I use for the filename that holds the data on my computer. Why? I don't want to accidentally overwrite this precious "source" of data.

```
# scrub_df <- read.csv ('BlackStntsModel230902.csv', head = TRUE, sep
# = ',')
scrub_df <- readRDS("BlackStntsModel230902.rds")
str(scrub_df)

## Classes 'tbl_df', 'tbl' and 'data.frame': 69 obs. of 25 variables:
## $ ID      : int 1 2 3 4 5 6 7 8 9 10 ...
## $ iRace1   : num 3 3 3 3 1 3 3 3 1 0 ...
## ..- attr(*, "label")= Named chr "1 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "1_iRace"
## $ iRace2   : num 1 NA 1 1 NA NA 3 NA NA 0 ...
## ..- attr(*, "label")= Named chr "2 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "2_iRace"
## $ iRace3   : num 3 NA NA 3 NA NA NA NA NA 3 ...
## ..- attr(*, "label")= Named chr "3 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "3_iRace"
## $ iRace4   : num NA NA NA NA NA NA NA NA NA 3 ...
## ..- attr(*, "label")= Named chr "4 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "4_iRace"
## $ iRace5   : logi NA NA NA NA NA NA NA ...
## ..- attr(*, "label")= Named chr "5 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "5_iRace"
## $ iRace6   : logi NA NA NA NA NA NA NA ...
## ..- attr(*, "label")= Named chr "6 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "6_iRace"
## $ iRace7   : logi NA NA NA NA NA NA NA ...
## ..- attr(*, "label")= Named chr "7 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "7_iRace"
## $ iRace8   : logi NA NA NA NA NA NA NA ...
## ..- attr(*, "label")= Named chr "8 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "8_iRace"
## $ iRace9   : logi NA NA NA NA NA NA NA ...
## ..- attr(*, "label")= Named chr "9 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "9_iRace"
## $ iRace10  : logi NA NA NA NA NA NA NA ...
## ..- attr(*, "label")= Named chr "10 - From your perspective as a student, which of the following ...
## ...- attr(*, "names")= chr "10_iRace"
## $ cmBiMulti: num 0 0 0 2 5 15 0 0 0 7 ...
```

```

## ..- attr(*, "label")= Named chr "Regarding race, what proportion of students were from each race group"
## ... - attr(*, "names")= chr "Race_10"
## $ cmBlack : num 0 5 10 6 5 20 0 0 0 4 ...
## ..- attr(*, "label")= Named chr "Regarding race, what proportion of students were from each race group"
## ... - attr(*, "names")= chr "Race_1"
## $ cmNBPoC : num 39 10 30 19 10 30 40 5 30 13 ...
## ..- attr(*, "label")= Named chr "Regarding race, what proportion of students were from each race group"
## ... - attr(*, "names")= chr "Race_7"
## $ cmWhite : num 61 85 60 73 80 35 60 90 70 73 ...
## ..- attr(*, "label")= Named chr "Regarding race, what proportion of students were from each race group"
## ... - attr(*, "names")= chr "Race_8"
## $ cmUnsure : num 0 0 0 0 0 0 5 0 3 ...
## ..- attr(*, "label")= Named chr "Regarding race, what proportion of students were from each race group"
## ... - attr(*, "names")= chr "Race_2"
## $ Belong_1 : num 6 4 NA 5 4 5 6 7 6 3 ...
## ..- attr(*, "label")= Named chr "Please indicate the degree to which you agree with the following statement"
## ... - attr(*, "names")= chr "Belong_1"
## $ Belong_2 : num 6 4 3 3 4 6 6 7 6 3 ...
## ..- attr(*, "label")= Named chr "Please indicate the degree to which you agree with the following statement"
## ... - attr(*, "names")= chr "Belong_2"
## $ Belong_3 : num 7 6 NA 2 4 5 5 7 6 3 ...
## ..- attr(*, "label")= Named chr "Please indicate the degree to which you agree with the following statement"
## ... - attr(*, "names")= chr "Belong_3"
## $ Blst_1 : num 5 6 NA 2 6 5 5 5 5 3 ...
## ..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
## ... - attr(*, "names")= chr "Blst_1"
## $ Blst_2 : num 3 6 5 2 1 1 4 4 3 5 ...
## ..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
## ... - attr(*, "names")= chr "Blst_2"
## $ Blst_3 : num 5 2 2 2 1 1 4 3 1 2 ...
## ..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
## ... - attr(*, "names")= chr "Blst_3"
## $ Blst_4 : num 2 2 2 2 1 2 4 3 2 3 ...
## ..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
## ... - attr(*, "names")= chr "Blst_4"
## $ Blst_5 : num 2 4 NA 2 1 1 4 4 1 3 ...
## ..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
## ... - attr(*, "names")= chr "Blst_5"
## $ Blst_6 : num 2 1 2 2 1 2 4 3 2 3 ...
## ..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
## ... - attr(*, "names")= chr "Blst_6"
## - attr(*, "column_map")=Classes 'tbl_df', 'tbl' and 'data.frame': 182 obs. of 7 variables
##   ..$ qname      : chr [1:182] "StartDate" "EndDate" "Status" "Progress" ...
##   ..$ description: chr [1:182] "Start Date" "End Date" "Response Type" "Progress" ...
##   ..$ main       : chr [1:182] "Start Date" "End Date" "Response Type" "Progress" ...
##   ..$ sub        : chr [1:182] "" "" "" "" ...
##   ..$ ImportId   : chr [1:182] "startDate" "endDate" "status" "progress" ...
##   ..$ timeZone   : chr [1:182] "America/Los_Angeles" "America/Los_Angeles" NA NA ...

```

```
## ..$ choiceId : chr [1:182] NA NA NA NA ...
```

Let's think about how the variables in our model should be measured:

- DV: Campus Climate for Black Students (as perceived by the respondent)
  - mean score of the 6 items on that scale (higher scores indicate a climate characterized by hostility, nonresponsiveness, and stigma)
  - 1 item needs to be reverse-coded
  - this scale was adapted from the LGBT Campus Climate Scale [[Szymanski and Bissonette, 2020](#)]
- IV: Belonging
  - mean score for the 3 items on that scale (higher scores indicate a greater sense of belonging)
  - this scale is taken from the Sense of Belonging subscale from the Perceived Cohesion Scale [[Bollen and Hoyle, 1990](#)]
- Proportion of classmates who are Black
  - a single item
- Proportion of instructional staff who are BIPOC
  - must be calculated from each of the single items for each instructor

To summarize, the Campus Climate and Belonging scales are traditional in the sense that they have items that we sum. The variable representing proportion of classmates who are Black is a single item. The variable representing the proportion of instructional staff who are BIPOC must be calculated in a manner that takes into consideration the there may be multiple instructors. The survey allowed a respondent to name up to 10 instructors.

```
str(scrub_df$iRace1)
```

```
## num [1:69] 3 3 3 3 1 3 3 3 1 0 ...
## - attr(*, "label")= Named chr "1 - From your perspective as a student, which of the follow
##   ..- attr(*, "names")= chr "1_iRace"
```

Looking at the structure of our data, the iRace(1 thru 10) variables are in “int” or integer format. This means that they are represented as whole numbers. We need them to be represented as factors. R handles factors represented as words well. Therefore, let's use our codebook to reformat this variable as a an ordered factor, with words instead of numbers.

Qualtrics imports many of the categorical variables as numbers. R often reads them numerically (integers or numbers). If they are directly converted to factors, R will sometimes collapse across missing numbers. In this example, if there is a race that is not represented (e.g., 2 for BiMulti), when the numbers are changed to factors, R will assume they are ordered and there is a consecutive series of numbers (0,1,2,3,4). If a number in the sequence is missing (0,1,3,4) and labels are applied, it will collapse across the numbers and the labels you think are attached to each number are not.

Therefore, it is ESSENTIAL to check (again and again ad nauseum) to ensure that your variables are recoding in a manner you understand.

One way to avoid this is to use the code below to identify the levels and the labels. When they are in order, they align and don't "skip" numbers. To quadruple check our work, we will recode into a new variable "tRace#" for "teacher" Race.

```
scrub_df$tRace1 = factor(scrub_df$iRace1, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace2 = factor(scrub_df$iRace2, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace3 = factor(scrub_df$iRace3, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace4 = factor(scrub_df$iRace4, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace5 = factor(scrub_df$iRace5, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace6 = factor(scrub_df$iRace6, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace7 = factor(scrub_df$iRace7, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace8 = factor(scrub_df$iRace8, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace9 = factor(scrub_df$iRace9, levels = c(0, 1, 2, 3, 4), labels = c("Black",
    "nBpoc", "BiMulti", "White", "NotNotice"))
scrub_df$tRace10 = factor(scrub_df$iRace10, levels = c(0, 1, 2, 3, 4),
    labels = c("Black", "nBpoc", "BiMulti", "White", "NotNotice"))
```

Let's check the structure to see if they are factors.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## vforcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyrr    1.3.0
## v purrr    1.0.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (http://conflicted.r-lib.org/) to force all conflicts to be
```

```
glimpse(scrub_df)
```

```
## Rows: 69  
## Columns: 35
```

```

## $ ID      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ iRace1  <dbl> 3, 3, 3, 3, 1, 3, 3, 1, 0, 2, 1, 1, 1, 3, 3, 3, 1, 3, 3, ~
## $ iRace2  <dbl> 1, NA, 1, 1, NA, NA, 3, NA, NA, 0, NA, NA, 3, NA, 3, 3, NA, ~
## $ iRace3  <dbl> 3, NA, NA, 3, NA, NA, NA, NA, NA, 3, NA, NA, NA, NA, 3, 1, N~
## $ iRace4  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 3, NA, NA, NA, NA, NA, 3~
## $ iRace5  <lgl> NA, ~
## $ iRace6  <lgl> NA, ~
## $ iRace7  <lgl> NA, ~
## $ iRace8  <lgl> NA, ~
## $ iRace9  <lgl> NA, ~
## $ iRace10 <lgl> NA, ~
## $ cmBiMulti <dbl> 0, 0, 0, 2, 5, 15, 0, 0, 0, 7, 0, 0, 20, 0, 9, 12, 0, 6, 6, ~
## $ cmBlack   <dbl> 0, 5, 10, 6, 5, 20, 0, 0, 0, 4, 0, 7, 0, 6, 9, 1, 21, 5, 6, ~
## $ cmNBPoC   <dbl> 39, 10, 30, 19, 10, 30, 40, 5, 30, 13, 80, 19, 0, 19, 15, 22~
## $ cmWhite   <dbl> 61, 85, 60, 73, 80, 35, 60, 90, 70, 73, 10, 74, 80, 0, 67, 5~
## $ cmUnsure  <dbl> 0, 0, 0, 0, 0, 0, 5, 0, 3, 10, 0, 0, 75, 0, 14, 0, 5, 0, ~
## $ Belong_1  <dbl> 6, 4, NA, 5, 4, 5, 6, 7, 6, 3, 6, 6, 3, 4, 3, 3, 4, 5, 1, 2, ~
## $ Belong_2  <dbl> 6, 4, 3, 3, 4, 6, 6, 7, 6, 3, 6, 6, 5, 4, 3, 3, 4, 6, 1, 2, ~
## $ Belong_3  <dbl> 7, 6, NA, 2, 4, 5, 5, 7, 6, 3, 5, 6, 4, 4, 3, 2, 4, 5, 1, 1, ~
## $ Blst_1    <dbl> 5, 6, NA, 2, 6, 5, 5, 5, 5, 3, NA, 4, 5, 6, 3, 4, 6, 4, 4, 4~
## $ Blst_2    <dbl> 3, 6, 5, 2, 1, 1, 4, 4, 3, 5, NA, 5, 1, 1, 3, 2, 1, 2, 5, 3, ~
## $ Blst_3    <dbl> 5, 2, 2, 2, 1, 1, 4, 3, 1, 2, 2, 1, 1, 1, 3, 2, 6, 2, 2, 2, ~
## $ Blst_4    <dbl> 2, 2, 2, 2, 1, 2, 4, 3, 2, 3, NA, 4, 3, 1, 3, 2, 1, 3, 2, 1, ~
## $ Blst_5    <dbl> 2, 4, NA, 2, 1, 1, 4, 4, 1, 3, 2, 2, 1, 1, 3, 2, 1, 2, 2, 1, ~
## $ Blst_6    <dbl> 2, 1, 2, 2, 1, 2, 4, 3, 2, 3, NA, 2, 1, 1, 3, 2, 2, 3, 2, 1, ~
## $ tRace1   <fct> White, White, White, White, nBpoc, White, White, White, nBpo~
## $ tRace2   <fct> nBpoc, NA, nBpoc, nBpoc, NA, NA, White, NA, NA, Black, NA, N~
## $ tRace3   <fct> White, NA, NA, White, NA, NA, NA, NA, White, NA, NA, NA, ~
## $ tRace4   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, White, NA, NA, NA, NA, N~
## $ tRace5   <fct> NA, ~
## $ tRace6   <fct> NA, ~
## $ tRace7   <fct> NA, ~
## $ tRace8   <fct> NA, ~
## $ tRace9   <fct> NA, ~
## $ tRace10  <fct> NA, ~

```

Calculating the proportion of the BIPOC instructional staff could likely be accomplished a number of ways. My searching for solutions resulted in this. Hopefully it's a fair balance between intuitive and elegant coding. First, I created code that

- created a new variable (count.BIPOC) by
  - summing across the tRace1 through tRace10 variables,
  - assigning a count of “1” each time the factor value was Black, nBpoc, or BiMulti

```

scrub_df$count.BIPOC <- apply(scrub_df[c("tRace1", "tRace2", "tRace3",
                                         "tRace4", "tRace5", "tRace6", "tRace7", "tRace8", "tRace9", "tRace10")],
                                1, function(x) sum(x %in% c("Black", "nBpoc", "BiMulti")))

```

Next, I created a variable that counted the number of non-missing values across the tRace1 through tRace10 variables.

```
scrub_df$count.nMiss <- apply(scrub_df[c("tRace1", "tRace2", "tRace3",
  "tRace4", "tRace5", "tRace6", "tRace7", "tRace8", "tRace9", "tRace10")],
  1, function(x) sum(!is.na(x)))
```

Now to calculate the proportion of BIPOC instructional faculty for each case.

```
scrub_df$iBIPOC_pr = scrub_df$count.BIPOC/scrub_df$count.nMiss
```

## 2.5.2 Missing Data Analysis: Whole df and Item level

In understanding missingness across the dataset, I think it is important to analyze and manage it, iteratively. We will start with a view of the whole df-level missingness. Subsequently, and consistent with the available information analysis [AIA; Parent [2013]] approach, we will score the scales and then look again at missingness, using the new information to update our decisions about how to manage it.

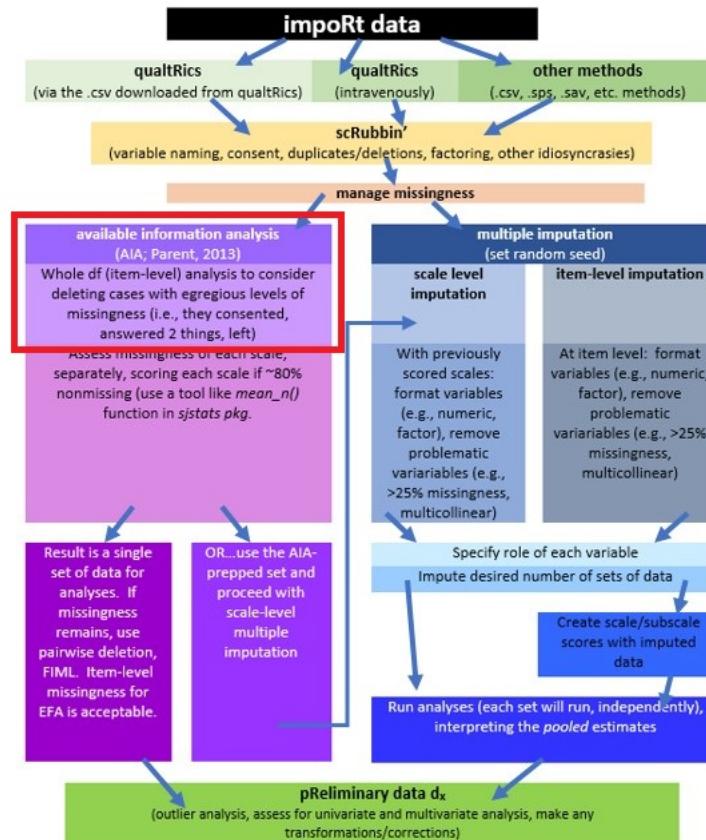


Figure 2.3: An image of our stage in the workflow for scrubbing and scoring data.

Because we just created a host of new variables in creating the *prop\_BIPOC* variable, let's downsize the df so that the calculations are sensible.

```
scrub_df <- (select(scrub_df, ID, iBIPOC_pr, cmBlack, Belong_1:Belong_3,
  Blst_1:Blst_6))
```

With a couple of calculations, we create a proportion of item-level missingness.

In this chunk I first calculate the number of missing (nmiss)

```
library(tidyverse)
#Calculating number and proportion of item-level missingness
scrub_df$nmiss <- scrub_df%>%
  select(iBIPOC_pr:Blst_6) %>% #the colon allows us to include all variables between the two
  is.na %>%
  rowSums

scrub_df<- scrub_df%>%
  dplyr::mutate(prop_miss = (nmiss/11)*100) #11 is the number of variables included in calcula
```

We can grab the descriptives for the *prop\_miss* variable to begin to understand our data. I will create an object from it so I can use it with inline

```
psych::describe(scrub_df$prop_miss)
```

```
##      vars   n  mean      sd median trimmed mad min max range skew kurtosis    se
## X1      1 69 7.77 22.61       0     1.59    0    0 100    100 3.04      8.19 2.72
```

#### CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%.

At the time that I am lecturing this, we do have some rather egregious missingness. At this point I will write code to eliminate cases with  $\geq 90\%$ .

```
scrub_df <- dplyr::filter(scrub_df, prop_miss <= 90) #update df to have only those with at le
```

To analyze missingness at this level, we need a df that has only the variables of interest. That is, variables like *ID* and the *prop\_miss* and *nmiss* variables we created will interfere with an accurate assessment of missingness. I will update our df to eliminate these.

```
# further update to exclude the n_miss and prop_miss variables
scrub_df <- scrub_df %>%
  dplyr::select(-c(ID, nmiss, prop_miss))
```

Missing data analysis commonly looks at proportions by:

- the entire df
- rows/cases/people

```
# what proportion of cells missing across entire dataset
formattable::percent(mean(is.na(scrub_df)))
```

```
## [1] 3.86%
```

```
# what proportion of cases (rows) are complete (nonmissing)
formattable::percent(mean(complete.cases(scrub_df)))
```

```
## [1] 87.88%
```

## CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

### 2.5.3 Analyzing Missing Data Patterns

One approach to analyzing missing data is to assess patterns of missingness.

Several R packages are popularly used for conducting such analyses. In the *mice* package, *md.pattern()* function provides a matrix with the number of columns + 1, in which each row corresponds to a missing data pattern (1 = observed, 0 = missing).

Rows and columns are sorted in increasing amounts of missing information.

The last column and row contain row and column counts, respectively.

```
mice_out <- mice::md.pattern(scrub_df, plot = TRUE, rotate.names = TRUE)
mice_out
write.csv(mice_out, file = "mice_out.csv") #optional to write it to a .csv file
```

The table lets us examine each missing pattern and see which variable(s) is/are missing. The output is in the form of a table that indicates the frequency of each pattern of missingness. Because I haven't (yet) figured out how to pipe objects from this table into the chapter, this text may differ from the patterns in the current data frame.

Each row in the table represents a different pattern of missingness. At the time of writing, there are 8 patterns of missing data. The patterns are listed in descending order of the least amount of missingness. The most common pattern (58 cases, top row) is one with no missing data. One case is missing one cell – one item assessing the campus climate for Black students, and so forth.

### 2.5.4 Can we identify the missing mechanisms?

To date, we do not have statistical tools that can accurately diagnose our patterns of missingness. You may have heard that “Little’s MCAR” is a helpful tool. Unfortunately, as Enders [2010] has noted, the tool is problematic. Perhaps the most significant one is that under the null hypothesis, a statistically significant test indicates that the missing data are MAR (missing at random) or MNAR (missing not at random); a non-significant test indicates the data are MCAR (missing completely at random) or MNAR. Consequently, regardless of the result, an MNAR circumstance cannot be ruled out. Correspondingly, the Little’s MCAR test has disappeared from the more reliable R packages that assess missingness.

Enders [2010] *Applied Missing Data Analysis* text does provide a set of [figures](#) (page 3) that illustrate common missing data patterns. Comparing these to the figure produced with *mice::mdpattern* our data looks somewhat monotonic – that is, as individuals completed the survey, they began to experience test fatigue and simply stopped responding. Diagnosing monotonicity requires that the variables in the dataset must be in the order in which the students completed them. If the variables have been re-ordered or if the surveys were presented to students in a randomized order, then more data manipulation would be required before attributing missingness to test fatigue.

Survey programs like Qualtrics offer the randomization of items within blocks (or blocks themselves). This can help distribute missingness caused by test fatigue so that more cases can be retained.

## 2.6 Scoring

So let’s get to work to score up the measures for our analysis. Each step of this should involve careful cross-checking with the [codebook](#).

### 2.6.1 Reverse scoring

As we discovered previously, in the scale that assesses campus climate (higher scores reflect a more negative climate) one of our items (Blst\_1, “My *institution* provides a supportive environment for Black students.”) requires reverse-coding.

To rescore:

- Create a *new* variable (this is essential) that is designated as the reversed item. We might put a the letter “r” (for reverse scoring) at the beginning or end: rBlst\_1 or Blst\_1r. It does not matter; just be consistent.
  - We don’t reverse score into the same variable because when you rerun the script, it just re-reverses the reversed score...into infinity. It’s very easy to lose your place.
- The reversal is an *equation* where you subtract the value in the item from the range/scaling + 1. For the our three items we subtract each item’s value from 8.

```
scrub_df <- scrub_df %>%
  dplyr::mutate(rBlst_1 = 8 - Blst_1) #if you had multiple items, you could add a pipe (%>%)
```

Per Parent [2013] we will analyze missingness for each scale, separately.

- We will calculate scale scores on each scale separately when 80% (roughly) of the data is present.
  - this is somewhat arbitrary, on 4 item scales, I would choose 75% (to allow one to be missing)
  - on the 3 item scale, I will allow one item to be missing (65%)
- After calculating the scale scores, we will return to analyzing the missingness, looking at the whole df

The `mean_n()` function of `sjstats` package has allows you to specify how many items (whole number) or what percentage of items should be present in order to get the mean. First, though, we should identify the variables (properly formatted, if rescoreing was needed) that should be included in the calculation of each scale and subscale.

In our case, the scale assessing belonging [Bollen and Hoyle, 1990, Hurtado and Carter, 1997] involves three items with no reversals. Our campus climate scale was adapted from Szymanski et al's LGBTQ College Campus Climate Scale [Szymanski and Bissonette, 2020]. While it has not been psychometrically evaluated for the purpose for which I am using it, I will follow the scoring structure in the journal article that introduces the measure. Specifically, the factor structure permits a total scale score and two subscales representing the college response and stigma.

```
# Making the list of variables
Belonging_vars <- c("Belong_1", "Belong_2", "Belong_3")
ResponseBL_vars <- c("rBlst_1", "Blst_4", "Blst_6")
StigmaBL_vars <- c("Blst_2", "Blst_3", "Blst_5")
ClimateBL_vars <- c("rBlst_1", "Blst_4", "Blst_6", "Blst_2", "Blst_3",
                     "Blst_5")

# Creating the new variables
scrub_df$Belonging <- sjstats::mean_n(scrub_df[, Belonging_vars], 0.65)
scrub_df$ResponseBL <- sjstats::mean_n(scrub_df[, ResponseBL_vars], 0.8)
scrub_df$StigmaBL <- sjstats::mean_n(scrub_df[, StigmaBL_vars], 0.8)
scrub_df$ClimateBL <- sjstats::mean_n(scrub_df[, ClimateBL_vars], 0.8)
```

Later it will be helpful to have a df with the item and scale-level variables. It will also be helpful if there is an ID for each case.

```
scrub_df <- scrub_df %>%
  dplyr::mutate(ID = row_number())

# moving the ID number to the first column; requires
scrub_df <- scrub_df %>%
  dplyr::select(ID, everything())
```

Let's save our `scrub_df` data for this and write it as an outfile. I will save it in both .rds and .csv formats so that you can use either one.

```
write.table(scrub_df, file = "BlStItmsScrs230902.csv", sep = ", ", col.names = TRUE,
            row.names = FALSE)
saveRDS(scrub_df, "BlStItmsScrs230902.rds")
```

## 2.7 Missing Analysis: Scale level

Let's return to analyzing the missingness, this time including the *scale level* variables (without the individual items) that will be in our statistical model(s).

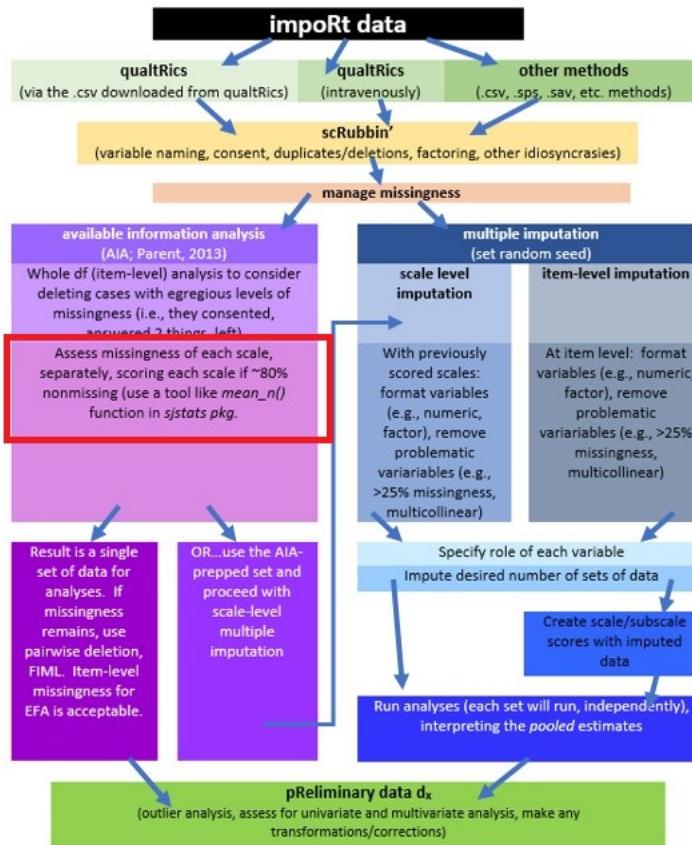


Figure 2.4: An image of our stage in the workflow for scrubbing and scoring data.

First let's get the df down to the variables we want to retain:

```
scored <- dplyr::select(scrub_df, iBIPOC_pr, cmBlack, Belonging, ResponseBL,
                        StigmaBL, ClimateBL)
ScoredCaseMiss <- nrow(scored) #I produced this object for the sole purpose of feeding the nu
ScoredCaseMiss
```

```
## [1] 66
```

Before we start our formal analysis of missingness at the scale level, let's continue to scrub by eliminating cases that will have too much missingness. In the script below we create a variable that counts the number of missing variables and then creates a proportion by dividing it by the number of total variables.

Using the *describe()* function from the *psych* package, we can investigate this variable.

```
# Create a variable (n_miss) that counts the number missing
scored$n_miss <- scored %>%
  dplyr::select(iBIPOC_pr:ClimateBL) %>%
  is.na %>%
  rowSums

# Create a proportion missing by dividing n_miss by the total number
# of variables (6) Pipe to sort in order of descending frequency to
# get a sense of the missingness
scored <- scored %>%
  dplyr::mutate(prop_miss = (n_miss/6) * 100) %>%
  arrange(desc(n_miss))

psych::describe(scored$prop_miss)

##      vars   n  mean      sd median trimmed mad min    max range skew kurtosis     se
## X1      1 66 3.79 12.33       0     0.31    0  0 66.67 66.67 3.44     11.77 1.52
```

#### CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

Across the 66 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%.

We need to decide what is our retention threshhold. Twenty percent seems to be a general rule of thumb. Let's delete all cases with missingness at 20% or greater.

```
# update df to have only those with at least 20% of complete data
# (this is an arbitrary decision)
scored <- dplyr::filter(scored, prop_miss <= 20)

# the variable selection just lops off the proportion missing
scored <- (select(scored, iBIPOC_pr:ClimateBL))

# this produces the number of cases retained
nrow(scored)
```

```
## [1] 61
```

CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

Across the 66 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 61 cases.

With a decision about the number of cases we are going to include, we can continue to analyze missingness.

## 2.8 Revisiting Missing Analysis at the Scale Level

We work with a df that includes only the variables in our model. In our case this is easy. In other cases (i.e., maybe there is an ID number) it might be good to create a subset just for this analysis.

Again, we look at missingness as the proportion of

- individual cells across the scored dataset, and
- rows/cases with nonmissing data

```
# percent missing across df
formattable::percent(mean(is.na(scored)))

## [1] 0.55%

# percent of rows with nonmissing data
formattable::percent(mean(complete.cases(scored)))

## [1] 96.72%
```

CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

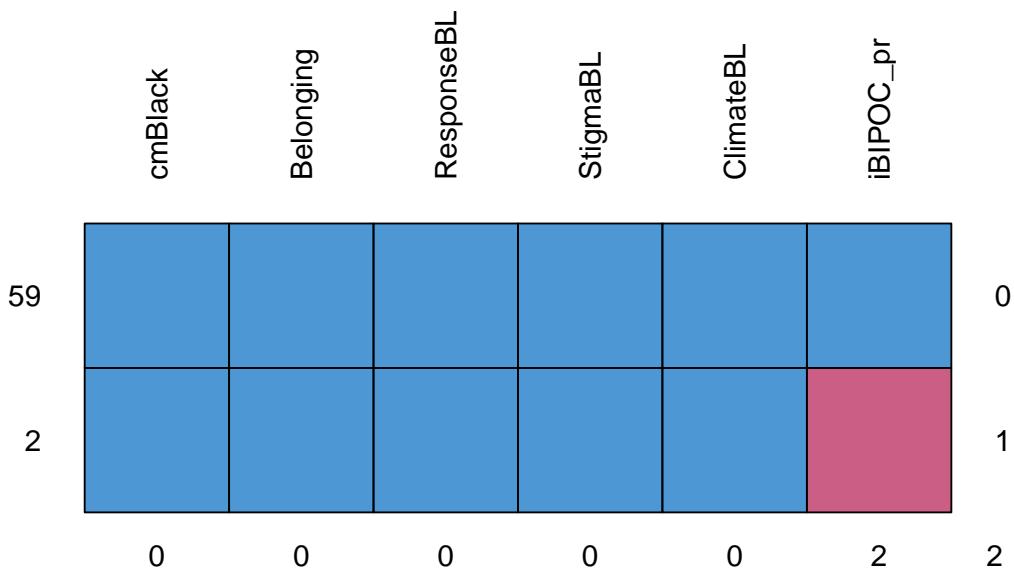
Across the 66 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 61 cases. In this dataset we had less than 1% (0.55%) missing across the df; 97% of the rows had nonmissing data.

Let's look again at missing patterns and mechanisms.

### 2.8.1 Scale Level: Patterns of Missing Data

Returning to the *mice* package, we can use the *md.pattern()* function to examine a matrix with the number of columns + 1 in which each row corresponds to a missing data pattern (1 = observed, 0 = missing). The rows and columns are sorted in increasing amounts of missing information. The last column and row contain row and column counts, respectively.

```
mice_ScaleLvl <- mice::md.pattern(scored, plot = TRUE, rotate.names = TRUE)
```



```
mice_ScaleLvl
```

```
##      cmBlack Belonging ResponseBL StigmaBL ClimateBL iBIPOC_pr
## 59       1        1         1        1        1        1 0
## 2        1        1         1        1        1        0 1
##       0        0         0        0        0        2 2
```

At the scale-level, this is much easier to interpret. There are 2 rows of data because there are only 2 patterns of missingness. The most common pattern is non-missing data ( $n = 59$ ).

If our statistical choice uses listwise deletion (i.e., the case is eliminated if one or more variables in the model has missing data), our sample size will be 59. As we will learn in later chapters, there are alternatives (i.e., specifying a FIML option in analyses that use maximum likelihood estimators) that can use all of the cases – even those with missing data.

## 2.8.2 R-ready for Analysis

At this stage the data is ready for analysis (data diagnostics). With the AIA approach [Parent, 2013] the following preliminary analyses would involve pairwise deletion (i.e., the row/case is dropped for that analysis, but included for all others):

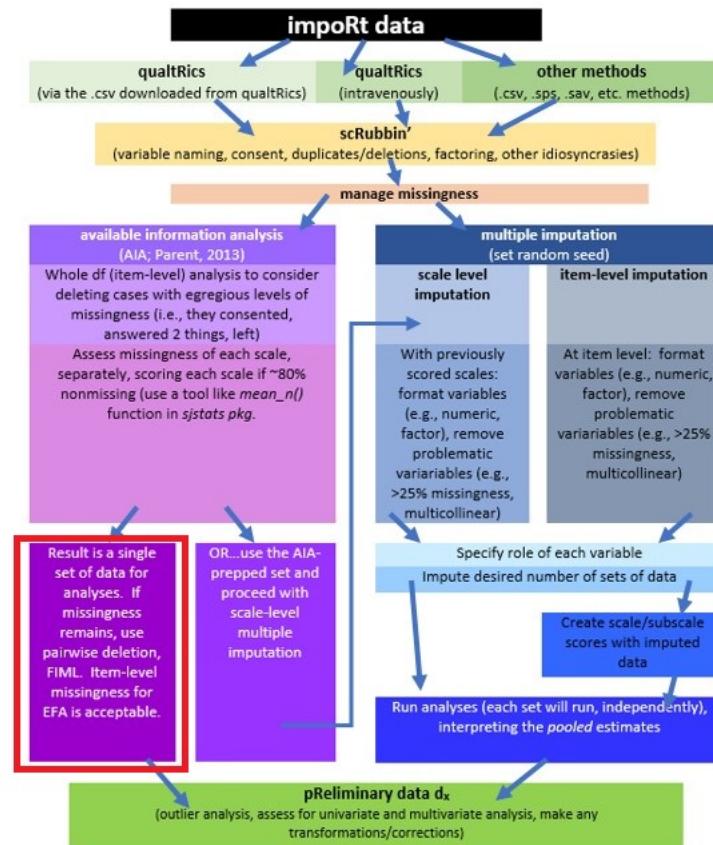


Figure 2.5: An image of our stage in the workflow for scrubbing and scoring data.

- data diagnostics
  - psychometric properties of scales, such as alpha coefficients
  - assessing assumptions such as univariate and multivariate normality, outliers, etc.
- preliminary analyses

- descriptives (means/standard deviations, frequencies)
- correlation matrices

AIA can also be used with primary analyses. Examples of how to manage missingness include:

- ANOVA/regression models
  - if completed with ordinary least squares, pairwise deletion would be utilized
- SEM/CFA models with observed, latent, or hybrid models
  - if FIML (we'll discuss later) is specified, all cases are used, even when there is missingness
- EFA models
  - these can handle item-level missingness
- Hierarchical linear modeling/multilevel modeling/mixed effects modeling
  - While all data needs to be present for a given cluster/wave, it is permissible to have varying numbers of clusters/waves per case

## 2.9 The APA Style Write-Up

## 2.10 Results

All analyses were completed in R Studio (v. RStudio 2023.06.1+524 “Mountain Hydrangea”) with R (v. 4.3.1).

### Missing Data Analysis and Treatment of Missing Data

Available item analysis (AIA; [Parent, 2013]) is a strategy for managing missing data that uses available data for analysis and excludes cases with missing data points only for analyses in which the data points would be directly involved. Parent (2013) suggested that AIA is equivalent to more complex methods (e.g., multiple imputation) across a number of variations of sample size, magnitude of associations among items, and degree of missingness. Thus, we utilized Parent's recommendations to guide our approach to managing missing data. Missing data analyses were conducted with tools in base R as well as the R packages, *psych* (v. 2.3.6) and *mice* (v. 3.16.0).

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 67%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data. At this stage in the analysis, we allowed all cases with less than 90% missing to continue to the scoring stage. Guided by Parent's [2013] AIA approach, scales with three items were scored if at least two items were non-missing; the scale with four items was scored if it at least three non-missing items; and the scale with six items was scored if it had at least five non-missing items.

Across the 66 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 61 cases. In this dataset we had less than 1% (0.55%) missing across the data set; 97% of the rows had nonmissing data.

## 2.11 Practice Problems

The three problems described below are designed to be continuations from the previous chapter (Scrubbing). You will likely encounter challenges that were not covered in this chapter. Search for and try out solutions, knowing that there are multiple paths through the analysis. The overall notion of the suggestions for practice are to (a) properly format three variables, (b) evaluate item-level missingness, (c) score any scales, (c) evaluate scale-level missingness, (d) provide an APA-style write-up, and (e) explain it to someone.

### 2.11.1 Problem #1: Reworking the Chapter Problem

If you chose this option in the prior chapter, you imported the data from Qualtrics, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, and wrote up the preliminary results.

### 2.11.2 Problem #2: Use the *Rate-a-Recent-Course* Survey, Choosing Different Variables

If you chose this option in the prior chapter, you chose a minimum of three variables from the *Rate-a-Recent-Course* survey to include in a simple statistical model. You imported the dat from Qualtrics, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest and wrote up the preliminary results.

### 2.11.3 Problem #3: Other data

If you chose this option in the prior chapter, you used raw data that was available to you. You imported it into R, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, and wrote up the preliminary results.

### 2.11.4 Grading Rubric

Assignment Component	Points Possible	Points Earned
1. Proper formatting of the items(s) in your first predictor variable	5	_____
2. Proper formatting of the items(s) in your second predictor variable	5	_____
3. Proper formatting of the items(s) your third predictor variable	5	_____
4. Proper formatting of your dependent variable	5	_____
4. Evaluate and interpret item-level missingness	5	_____

Assignment Component	Points Possible	Points Earned
5. Score any scales/subscales	5	_____
7. Represent your work in an APA-style write-up (added to the writeup in the previous chapter)	5	_____
8. Explanation to grader	5	_____
<b>Totals</b>	<b>45</b>	_____

A *homeworked example* for the Scrubbing, Scoring, and DataDx lessons (combined) follows the **Data Dx** lesson.

# Chapter 3

## Data Dx

### [Screencasted Lecture Link](#)

The focus of this chapter is *data diagnostics*. We are asking the question, “Does the data have the appropriate characteristics for the analysis we want to perform?” Some statistics are more robust than others to violations of the assumptions about the characteristics of the data. None-the-less, we must report these characteristics when we disseminate the results.

### 3.1 Navigating this Lesson

There is about 45 minutes of lecture. If you work through the materials with me it would be plan for an additional hour.

While the majority of R objects and data you will need are created within the R script that sources the chapter, there are a few that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER’s [introduction](#)

#### 3.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Conduct and interpret critical data diagnostics, including
  - alpha coefficients
  - skew
  - kurtosis
- Assess univariate and multivariate normality
- Identify options for managing outliers and skewed data
- Articulate a workflow for data preparation, including scrubbing, scoring, and data diagnostics

### 3.1.2 Planning for Practice

The suggestions from practice are a continuation from the two prior chapters. If you have completed one or more of those assignments, you should have started with a raw dataset and then scrubbed and scored it. This chapter will involve running basic data diagnostics. Options of graded complexity could include:

- Repeating the steps in the chapter with the most recent data from the Rate-A-Recent-Course survey; differences will be in the number of people who have completed the survey since the chapter was written.
- Use the dataset that is the source of the chapter, but score a different set of items that you choose.
- Begin with raw data to which you have access.

### 3.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Parent, M. C. (2013). Handling item-level missing data: Simpler is just as good. *The Counseling Psychologist*, 41(4), 568–600. <https://doi.org/10.1177/001100012445176>
  - The purpose of Parent’s article was to argue that complex and resource-intensive procedures like multiple imputation are unnecessary. Following a simulation that supports his claims, Parent provides some guidelines to follow for the AIA approach.
- Kline, R. B. (2015). Data preparation and psychometrics review. In *Principles and Practice of Structural Equation Modeling*, Fourth Edition. Guilford Publications. <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4000663>
  - Kline’s chapter is my “go-to” for making decisions about preparing data for analysis.

### 3.1.4 Packages

The packages used in this lesson are embedded in this code. When the hashtags are removed, the script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# if(!require(tidyverse)){install.packages('tidyverse')} #this  
# includes dplyr if(!require(psych)){install.packages('psych')}  
# if(!require(apaTables)){install.packages('apaTables')}
```

## 3.2 Workflow for Preliminary Data Diagnostics

The same workflow guides us through the Scrubbing, Scoring, and Data Dx chapters. At this stage we have

- imported our raw data from Qualtrics,
- scrubbed the data by applying our inclusion and exclusion criteria, and
- used Parent's available information approach [AIA; -Parent [2013]] for determining the acceptable amount of missingness for each scale, and
- prepared variables and scored them.

We are now ready to engage in data diagnostics for the statistical model we will test.

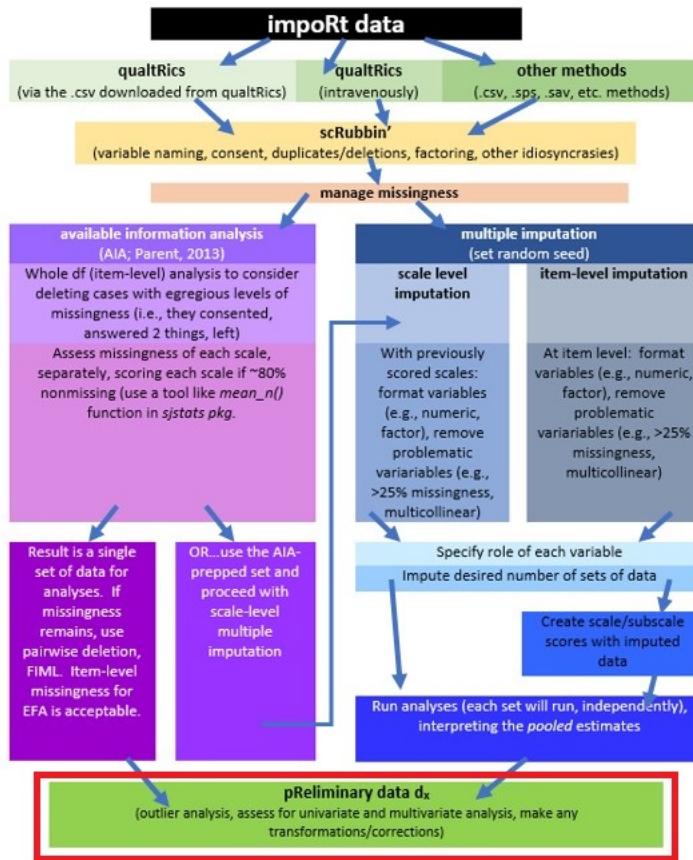


Figure 3.1: An image of our stage in the workflow for scrubbing and scoring data.

### 3.3 Research Vignette

The research vignette comes from the survey titled, [Rate-a-Recent-Course: A ReCentering Psych Stats Exercise](#) and is explained in the [scrubbing chapter](#). In the scoring chapter we prepared four variables for analysis. Details for these are in our [codebook](#).

Variable recap:

- Perceived Campus Climate for Black Students includes 6 items, one of which was reverse scored. This scale was adapted from Szymanski et al.'s [2020] Campus Climate for LGBTQ students. It has not been evaluated for use with other groups. The Szymanski et al. analysis

suggested that it could be used as a total scale score, or divided into three items each that assess

- College response to LGBTQ students (items 6, 4, 1)
- LGBTQ stigma (items 3, 2, 5)
- Sense of Belonging includes 3 items. This is a subscale from Bollen and Hoyle's [1990] Perceived Cohesion Scale. There are no items on this scale that require reversing.
- Percent of Black classmates is a single item that asked respondents to estimate the proportion of students in various racial categories
- Percent of BIPOC instructional staff, similarly, asked respondents to identify the racial category of each member of their instructional staff

As we noted in the [scrubbing chapter](#), our design has notable limitations. Briefly, (a) owing to the open source aspect of the data we do not ask about the demographic characteristics of the respondent; (b) the items that ask respondents to *guess* the identities of the instructional staff and to place them in broad categories, (c) we do not provide a “write-in” a response. We made these decisions after extensive conversation with stakeholders. The primary reason for these decisions was to prevent potential harm (a) to respondents who could be identified if/when the revealed private information in this open-source survey, and (b) trolls who would write inappropriate or harmful comments.

As I think about “how these variables go together” (which is often where I start in planning a study), imagine a parallel mediation. That is the perception of campus climate for Black students would be predicted by the respondent’s sense of belonging, mediated in separate paths through the proportion of classmates who are Black and the proportion of BIPOC instructional staff.

*I would like to assess the model by having the instructional staff variable to be the %Black instructional staff. At the time that this lecture is being prepared, there is not sufficient Black representation in the staff to model this.*

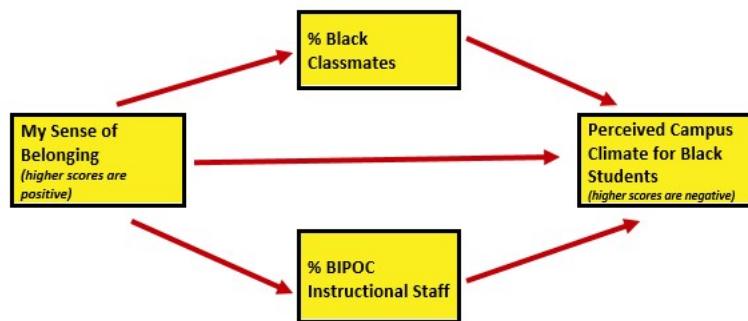


Figure 3.2: An image of the statistical model for which we are preparing data.

I will finish up this chapter by conducting a regression. Because parallel mediation can be complicated (I teach it in a later chapter), I will demonstrate use of our prepared variables with a simple multiple regression.

First, though, let’s take a more conceptual look at issues regarding missing data. We’ll come back to details of the survey as we work with it.

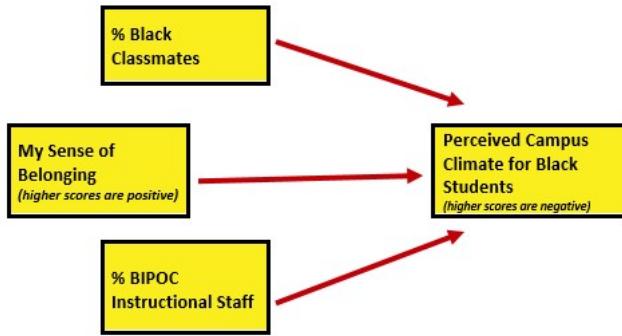


Figure 3.3: An image of the statistical model for which we are preparing data.

### 3.4 Internal Consistency of Scales/Subscales

Alpha coefficients are *reliability coefficients* that assess the *internal consistency* of an instrument. It asks, “For each person, are responses *consistently* high, or medium, or low?” To the degree that they are (meaning there are high inter-item correlations), the internal consistency coefficient will be high. We want values  $>.80$ . There are numerous problems with alpha coefficients. The biggest one is that they are influenced by sample size – longer scales have higher alpha coefficients [Cortina, 1993]. Fourteen seems to be a magic number where we begin to not trust the high alpha coefficient. I address this more thoroughly – offering an alternative – in psychometrics. While there is much criticism about the usefulness of the alpha coefficient [Sijtsma, 2009], researchers continue to use the alpha coefficient as an indicator of the internal consistency of scales that consist of multiple items and contain several variables.

We need item level data to compute an alpha coefficient. The easiest way to get an alpha coefficient is to feed the *alpha()* function (*psych* package) a concatenated list of items (with any items already reverse-scored). There should be no extra items. In the [scoring chapter](#) we already reverse-coded the single item in the campus climate scale, so we are ready to calculate alphas.

The df from which I am pulling data was created and written as an outfile in the [scoring chapter](#). You may also download the file from the [Github site](#) that hosts the chapter. Be sure to place the file in the same folder as the .rmd file. This particular df has item-level data. I am working with the .rds file. In case this is problematic for you, I have also provided code to import a .csv version of the file.

```
item_scores_df <- readRDS("BlStItmsScrs230902.rds")
# item_scores_df <- read.csv('BlStItmsScrs230902.csv', header = TRUE)
```

Within the *psych::alpha* function we can retrieve alpha coefficients for the specific variables of interest by imbedding a concatenated list. A priori, we are planning to use the campus climate scale as a total score. However, we'll go ahead and also calculate alpha coefficients for the subscales because (a) it's good practice and (b) if the alpha is low, a *reason* might show up in one of the subscales.

```
# alpha for the belonging scale
psych::alpha(item_scores_df[c("Belong_1", "Belong_2", "Belong_3")])
```

Reliability analysis  
Call: psych::alpha(x = item\_scores\_df[c("Belong\_1", "Belong\_2", "Belong\_3")])

	raw_alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd	median_r
	0.95	0.95	0.93	0.87	21	0.0099	4	1.5	0.88

95% confidence boundaries  
lower alpha upper  
Feldt 0.93 0.95 0.97  
Duhachek 0.93 0.95 0.97

Reliability if an item is dropped:

	raw_alpha	std.alpha	G6(smc)	average_r	S/N	alpha	se	var.r	med.r
Belong_1	0.94	0.94	0.88	0.88	15	0.016	NA	0.88	
Belong_2	0.92	0.92	0.85	0.85	11	0.020	NA	0.85	
Belong_3	0.94	0.94	0.89	0.89	16	0.015	NA	0.89	

Item statistics  
n raw.r std.r r.cor r.drop mean sd  
Belong\_1 64 0.95 0.95 0.92 0.90 4.1 1.5  
Belong\_2 65 0.96 0.96 0.94 0.92 4.1 1.6  
Belong\_3 64 0.95 0.95 0.91 0.89 3.8 1.5

Non missing response frequency for each item  
1 2 3 4 5 6 7 miss  
Belong\_1 0.02 0.14 0.23 0.17 0.22 0.17 0.05 0.03  
Belong\_2 0.03 0.14 0.22 0.22 0.15 0.20 0.05 0.02  
Belong\_3 0.05 0.19 0.19 0.23 0.20 0.09 0.05 0.03

For each scale I will capture a statement for the APA style write-up. Because these values are typically reported with each measure (and not in the preliminary results), I won't create a cumulative write-up.

Cronbach's alpha for the belonging scale was 0.95.

```
# alpha for the campus climate for Black students scale
psych::alpha(item_scores_df[c("rBlst_1", "Blst_2", "Blst_3", "Blst_4",
"Blst_5", "Blst_6")])
```

Reliability analysis  
Call: psych::alpha(x = item\_scores\_df[c("rBlst\_1", "Blst\_2", "Blst\_3",

```

"Blst_4", "Blst_5", "Blst_6")])

raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd median_r
  0.85      0.87     0.87      0.52 6.5 0.03  2.5 1.1      0.52

95% confidence boundaries
  lower alpha upper
Feldt    0.78  0.85  0.90
Duhachek 0.79  0.85  0.91

Reliability if an item is dropped:
  raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
rBlst_1    0.85      0.87     0.87      0.57 6.5 0.031 0.029  0.57
Blst_2     0.87      0.88     0.87      0.59 7.1 0.026 0.019  0.56
Blst_3     0.83      0.85     0.85      0.54 5.8 0.034 0.029  0.50
Blst_4     0.80      0.82     0.82      0.48 4.6 0.041 0.027  0.48
Blst_5     0.79      0.81     0.81      0.46 4.3 0.042 0.024  0.47
Blst_6     0.80      0.82     0.81      0.48 4.6 0.040 0.021  0.50

Item statistics
  n raw.r std.r r.cor r.drop mean   sd
rBlst_1 60 0.69 0.67 0.56 0.52 3.4 1.6
Blst_2 64 0.68 0.62 0.51 0.46 3.0 1.8
Blst_3 63 0.71 0.74 0.66 0.59 2.0 1.2
Blst_4 62 0.85 0.86 0.84 0.77 2.5 1.3
Blst_5 63 0.89 0.89 0.89 0.82 2.0 1.2
Blst_6 63 0.83 0.86 0.86 0.77 2.1 1.3

Non missing response frequency for each item
  1   2   3   4   5   6   7 miss
rBlst_1 0.10 0.23 0.20 0.25 0.08 0.10 0.03 0.09
Blst_2  0.33 0.16 0.09 0.17 0.16 0.06 0.03 0.03
Blst_3  0.44 0.33 0.06 0.11 0.03 0.02 0.00 0.05
Blst_4  0.27 0.34 0.15 0.18 0.05 0.00 0.02 0.06
Blst_5  0.46 0.30 0.05 0.14 0.05 0.00 0.00 0.05
Blst_6  0.38 0.35 0.11 0.08 0.06 0.02 0.00 0.05

```

Cronbach's alpha for the campus climate scale was 0.87.

Since this value is  $\geq .80$ , it is within the realm of acceptability. Let's go ahead, though, and examine its subscales.

```

# alpha for the stigma scale of the campus climate for Black students
# scale
psych::alpha(item_scores_df[c("Blst_3", "Blst_2", "Blst_5")])

```

```

Reliability analysis
Call: psych::alpha(x = item_scores_df[c("Blst_3", "Blst_2", "Blst_5")])

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd median_r
  0.69      0.73      0.69      0.47 2.7 0.065  2.3 1.2      0.54

  95% confidence boundaries
    lower alpha upper
Feldt     0.54  0.69  0.80
Duhachek  0.57  0.69  0.82

Reliability if an item is dropped:
  raw_alpha std.alpha G6(smc) average_r  S/N alpha se var.r med.r
Blst_3     0.67      0.70      0.54      0.54 2.35  0.074   NA  0.54
Blst_2     0.75      0.75      0.60      0.60 3.03  0.061   NA  0.60
Blst_5     0.41      0.43      0.28      0.28 0.76  0.135   NA  0.28

Item statistics
  n raw.r std.r r.cor r.drop mean   sd
Blst_3 63  0.72  0.78  0.62  0.46   2 1.2
Blst_2 64  0.82  0.75  0.55  0.46   3 1.8
Blst_5 63  0.87  0.89  0.83  0.70   2 1.2

Non missing response frequency for each item
  1   2   3   4   5   6   7 miss
Blst_3 0.44 0.33 0.06 0.11 0.03 0.02 0.00 0.05
Blst_2 0.33 0.16 0.09 0.17 0.16 0.06 0.03 0.03
Blst_5 0.46 0.30 0.05 0.14 0.05 0.00 0.00 0.05

```

Cronbach's alpha for the campus climate stigma subscale was 0.73.

```

# alpha for the campus responsiveness scale of the campus climate for
# Black students scale
psych::alpha(item_scores_df[c("rBlst_1", "Blst_4", "Blst_6")])

```

```

Reliability analysis
Call: psych::alpha(x = item_scores_df[c("rBlst_1", "Blst_4", "Blst_6")])

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd median_r
  0.79      0.81      0.76      0.58 4.2 0.045  2.7 1.2      0.52

  95% confidence boundaries
    lower alpha upper
Feldt     0.69  0.79  0.87
Duhachek  0.71  0.79  0.88

```

```

Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
rBlst_1      0.86      0.86     0.75      0.75 6.0     0.035    NA   0.75
Blst_4       0.64      0.65     0.48      0.48 1.8     0.087    NA   0.48
Blst_6       0.68      0.68     0.52      0.52 2.1     0.078    NA   0.52

Item statistics
    n raw.r std.r r.cor r.drop mean   sd
rBlst_1 60  0.81  0.78  0.58  0.53  3.4 1.6
Blst_4   62  0.88  0.89  0.84  0.72  2.5 1.3
Blst_6   63  0.85  0.87  0.81  0.69  2.1 1.3

Non missing response frequency for each item
    1   2   3   4   5   6   7 miss
rBlst_1 0.10 0.23 0.20 0.25 0.08 0.10 0.03 0.09
Blst_4   0.27 0.34 0.15 0.18 0.05 0.00 0.02 0.06
Blst_6   0.38 0.35 0.11 0.08 0.06 0.02 0.00 0.05

```

Cronbach's alpha for the campus climate responsiveness subscale was 0.80. Between the two subscales, it looks as if the responsiveness subscale is more internally consistent.

## 3.5 Distributional Characteristics of the Variables

### 3.5.1 Evaluating Univariate Normality

Statistics like ANOVA and regression each have a set of assumptions about the distributional characteristics of the data. In most of the chapters in this OER we review those assumptions and how to evaluate them. Common across many statistics is the requirement of univariate and multivariate normality. Let's take a look at the variables we will use in our analysis and assess those.

We can continue to work from the df we uploaded at the beginning of the chapter to do this work. Let's take a quick peek. This df has the item-level data (we used it for the alpha coefficients); the scale and subscale scores; and the two items that assess proportion of instructional staff that are BIPOC and proportion of classmates that are BIPOC.

The `str()` function let's us look at the variable format/measurement level of each variable.

```
str(item_scores_df)
```

```

Classes 'tbl_df', 'tbl' and 'data.frame': 66 obs. of 17 variables:
 $ ID        : int 1 2 3 4 5 6 7 8 9 10 ...
 $ iBIPOC_pr : num 0.333 0 0.5 0.333 1 ...
 $ cmBlack   : num 0 5 10 6 5 20 0 0 0 4 ...
 ..- attr(*, "label")= Named chr "Regarding race, what proportion of students were from each"
 ...- attr(*, "names")= chr "Race_1"
```

```

$ Belong_1 : num  6 4 NA 5 4 5 6 7 6 3 ...
..- attr(*, "label")= Named chr "Please indicate the degree to which you agree with the following statement"
... ..- attr(*, "names")= chr "Belong_1"
$ Belong_2 : num  6 4 3 3 4 6 6 7 6 3 ...
..- attr(*, "label")= Named chr "Please indicate the degree to which you agree with the following statement"
... ..- attr(*, "names")= chr "Belong_2"
$ Belong_3 : num  7 6 NA 2 4 5 5 7 6 3 ...
..- attr(*, "label")= Named chr "Please indicate the degree to which you agree with the following statement"
... ..- attr(*, "names")= chr "Belong_3"
$ Blst_1   : num  5 6 NA 2 6 5 5 5 5 3 ...
..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
... ..- attr(*, "names")= chr "Blst_1"
$ Blst_2   : num  3 6 5 2 1 1 4 4 3 5 ...
..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
... ..- attr(*, "names")= chr "Blst_2"
$ Blst_3   : num  5 2 2 2 1 1 4 3 1 2 ...
..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
... ..- attr(*, "names")= chr "Blst_3"
$ Blst_4   : num  2 2 2 2 1 2 4 3 2 3 ...
..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
... ..- attr(*, "names")= chr "Blst_4"
$ Blst_5   : num  2 4 NA 2 1 1 4 4 1 3 ...
..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
... ..- attr(*, "names")= chr "Blst_5"
$ Blst_6   : num  2 1 2 2 1 2 4 3 2 3 ...
..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
... ..- attr(*, "names")= chr "Blst_6"
$ rBlst_1  : num  3 2 NA 6 2 3 3 3 3 5 ...
..- attr(*, "label")= Named chr "Each item below asks you to rate elements of campus climate"
... ..- attr(*, "names")= chr "Blst_1"
$ Belonging : num  6.33 4.67 NA 3.33 4 5.33 5.67 7 6 3 ...
$ ResponseBL: num  2.33 1.67 2 3.33 1.33 2.33 3.67 3 2.33 3.67 ...
$ StigmaBL  : num  3.33 4 3.5 2 1 1 4 3.67 1.67 3.33 ...
$ ClimateBL : num  2.83 2.83 NA 2.67 1.17 1.67 3.83 3.33 2 3.5 ...
- attr(*, "column_map")=Classes 'tbl_df', 'tbl' and 'data.frame': 182 obs. of 7 variables:
..$ qname      : chr [1:182] "StartDate" "EndDate" "Status" "Progress" ...
..$ description: chr [1:182] "Start Date" "End Date" "Response Type" "Progress" ...
..$ main       : chr [1:182] "Start Date" "End Date" "Response Type" "Progress" ...
..$ sub        : chr [1:182] "" "" "" ...
..$ ImportId   : chr [1:182] "startDate" "endDate" "status" "progress" ...
..$ timeZone   : chr [1:182] "America/Los_Angeles" "America/Los_Angeles" NA NA ...
..$ choiceId   : chr [1:182] NA NA NA NA ...

```

The difference between “int” (integer) and “num” (numerical) is that integers are limited to whole numbers. For the statistics used in this lesson, both are acceptable formats for the variables.

```
# the script may look a little complicated; I could have simply
# written: describe(item_scores_df) because I only wanted only a few
# variables, I provided them in a concatenated: list [c('iBIPOC_pr',
# 'cmBlack', 'Belonging', 'ClimateBL')] I used type =1 so that we can
# interpret skew and kurtosis along Kline's recommendations I created
# an object from the descriptive results, this can be used to export
# the results for easier table making or manipulation outside of R

descriptives <- psych::describe(item_scores_df[c("iBIPOC_pr", "cmBlack",
    "Belonging", "ClimateBL")], type = 1)
# When we capture results in an object, we need to write it below so
# the results will display
descriptives
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
iBIPOC_pr	1	64	0.35	0.39	0.25	0.32	0.37	0	1.00	1.00	0.64	-1.05
cmBlack	2	66	8.20	8.02	5.50	7.24	8.15	0	30.00	30.00	0.95	0.05
Belonging	3	64	4.03	1.47	4.00	4.03	1.48	1	7.00	6.00	0.03	-0.76
ClimateBL	4	61	2.48	1.09	2.33	2.41	0.99	1	5.67	4.67	0.56	0.04
		se										
iBIPOC_pr		0.05										
cmBlack		0.99										
Belonging		0.18										
ClimateBL		0.14										

```
# this can be useful if you wish to manually format the data for an
# APA style table
write.csv(descriptives, file = "DataDx_descripts.csv")
```

Skew and kurtosis are one way to evaluate whether or not data are normally distributed. When we use the “type=1” argument, the skew and kurtosis indices in the *psych* package can be interpreted according to Kline’s [2016a] guidelines. Regarding skew, values greater than the absolute value of 3.0 are generally considered “severely skewed.” Regarding kurtosis, “severely kurtotic” is argued to be anywhere greater 8 to 20. Kline recommended using a conservative threshold of the absolute value of 10. The skew and kurtosis values for our variables fall well below these thresholds.

We can also apply the Shapiro-Wilk test of normality to each of our variables. When the *p* value is < .05, the variable’s distribution is deviates from a normal distribution to a degree that is statistically significant. Below, the plotting of the histogram with a normal curve superimposed shows how the distribution approximates one that is normal.

```
# The shapiro-test is in base R; it's specification is simple:
# shapiro.test(df$variable) I added the object (and had to list it
# below) so I can use the inline text function
shapiro.test(item_scores_df$cmBlack)
```

```
Shapiro-Wilk normality test

data: item_scores_df$cmBlack
W = 0.87796, p-value = 0.000009899

shapiro.test(item_scores_df$iBIPOC_pr)
```

```
Shapiro-Wilk normality test

data: item_scores_df$iBIPOC_pr
W = 0.78725, p-value = 0.00000003181

shapiro.test(item_scores_df$Belonging)
```

```
Shapiro-Wilk normality test

data: item_scores_df$Belonging
W = 0.97262, p-value = 0.1654

shapiro.test(item_scores_df$ClimateBL)
```

```
Shapiro-Wilk normality test

data: item_scores_df$ClimateBL
W = 0.95102, p-value = 0.01613
```

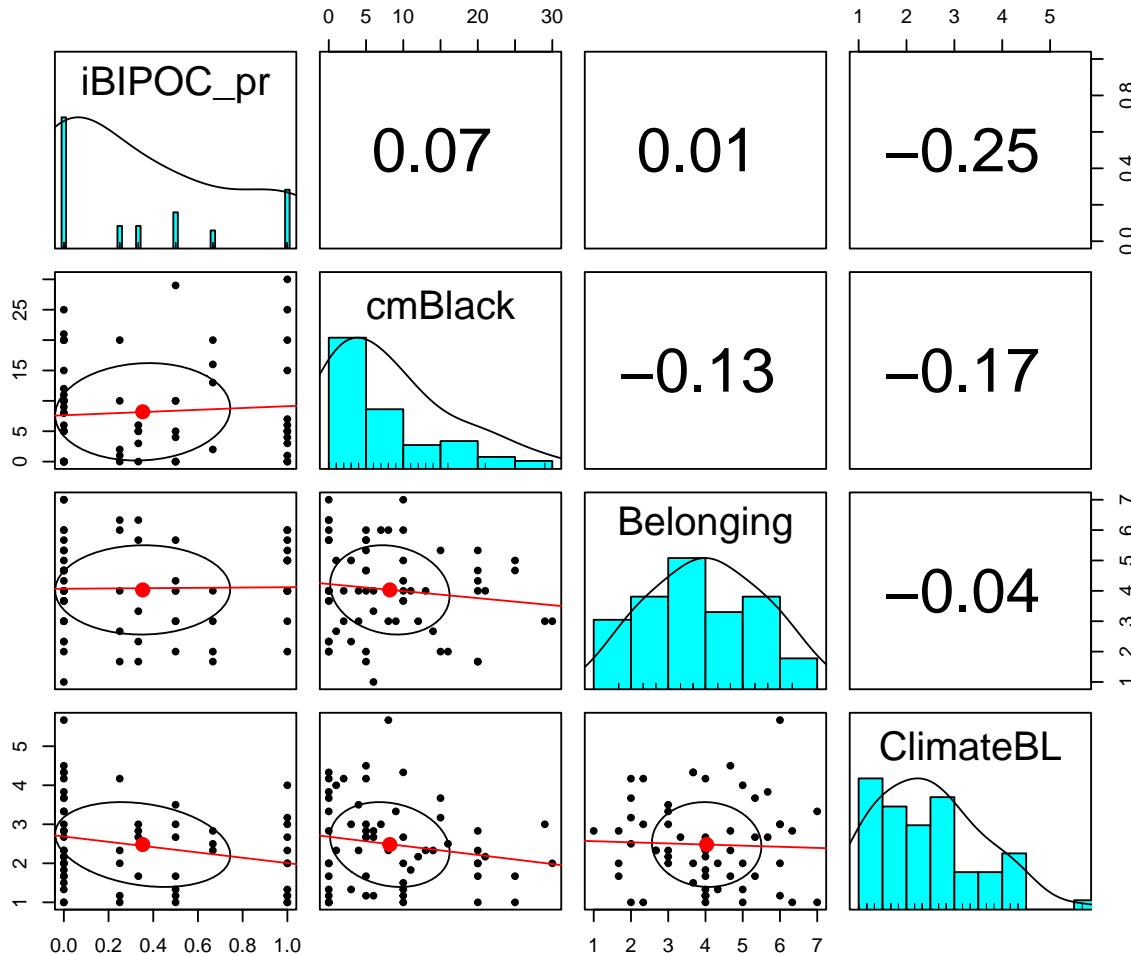
### 3.5.2 Pairs Panels

As we work our way from univariate to multivariate inspection of our data, let's take a look at the bivariate relations.

The *pairs.panels()* function from the *psych* package is useful for showing the relationship between variables (probably no more than 10) in a model.

- The lower half is a scatterplot between the two variables with a regression line (red) and mean (dot).
- The diagonal is a histogram of each variable.
- The upper half of is the correlation coefficient between the two variables.

```
psych::pairs.panels(item_scores_df[c("iBIPOC_pr", "cmBlack", "Belonging",
  "ClimateBL")], stars = TRUE, lm = TRUE)
```



The histograms displayed in the diagonal graph for us what we learned from the Shapiro Wilk's test of normality. We can clearly see the non-normal distribution in the iBIPOC\_pr and cmBlack variables.

#### CUMULATIVE CAPTURE FOR THE APA STYLE WRITE-UP:

Regarding the distributional characteristics of the data, skew and kurtosis values of the variables fell below the values of 3 (skew) and 10 (kurtosis) that Kline suggests are concerning [2016b]. Results of the Shapiro-Wilk test of normality indicate that our variables assessing the proportion of classmates who are Black ( $W = 0.878, p < 0.001$ ) and the proportion of BIPOC instructional staff ( $W = 0.787, p < 0.001$ ) are statistically significantly different than a normal distribution. Similarly the scale assessing the respondent's perception of campus climate for Black students ( $W = 0.951, p = 0.016$ ) differed significantly from a normal distribution. In all three cases the skew

values and histograms suggested a somewhat positive skew. That is, there were predominantly low proportions of instructional staff who are BIPOC and classmates who are Black, and the perceptions of campus climate for Black students was evaluated somewhat favorably. The scales assessing the respondent's belonging ( $0.973, p = 0.165$ ) did not differ significantly from a normal distribution.

What would we do in the case of a univariate outlier? I find Kline's [2016b] chapter on data preparation and management to be extremely useful. He provides ideas for more complex analysis of both univariate and multivariate normality and provides suggestions that range from recoding an extreme value to the next most extreme that is within three standard deviations of the mean to more complicated transformations. First, though we need to further examine the relationships between variables. We do that, next.

## 3.6 Evaluating Multivariate Normality

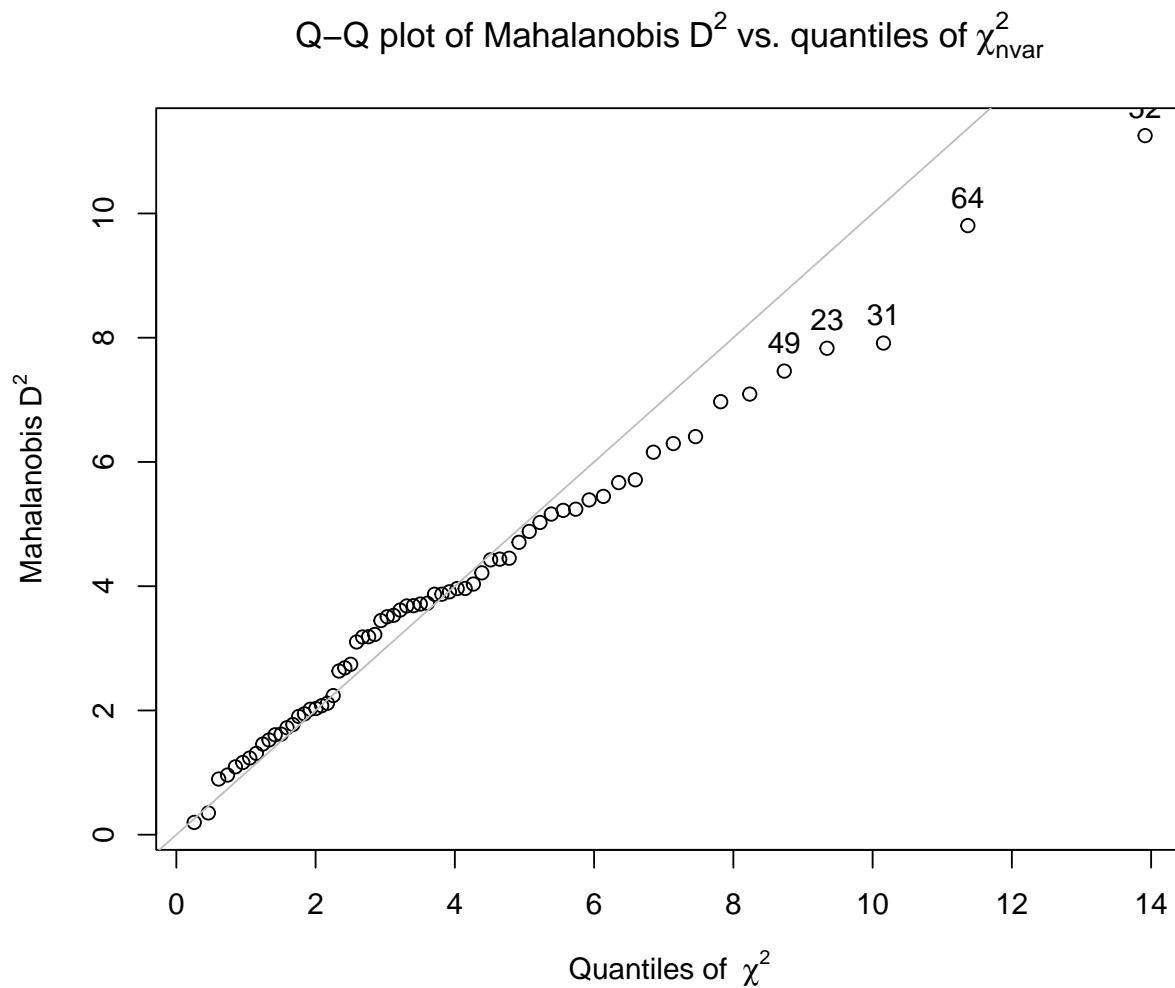
**Multivariate outliers** have extreme scores on two or more variables, or a pattern of scores that is atypical. For example, a case may have scores between two and three standard deviations above the mean on all variables, even though no case would be extreme. A common method of multivariate outlier detection is the **Mahalanobis distance** ( $D_M^2$ ). This indicates the distance in variance units between the profile of scores for that case and the vector of sample means, or **centroid**, correcting for intercorrelations.

The *outlier()* function from the *psych* package tells us how far each datapoint is from the multivariate centroid of the data. That is, find the squared Mahalanobis distance for each data point and compare it to the expected values of  $\chi^2$ . The *outlier()* protocol also produces a Q-Q (quantile-quantile) plot with the  $n$  most extreme data points labeled.

The code below appends the Mahalanobis values to the dataframe. It is easy, then, to identify, sort, and examine the most extreme values (relative to the rest of the data in their case/row) to make decisions about their retention or adjustment.

Numeric variables are required in the of the calculation of the Mahalanobis.

```
item_scores_df$Mahal <- psych::outlier(item_scores_df[c("iBIPOC_pr", "cmBlack",
  "Belonging", "ClimateBL")])
```



Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles (the number varies; you can see it on the X axis) calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution. To the degree that the plotted line stays on the straight line (representing the theoretical normal distribution), the data is multivariate normally distributed.

It is possible, then to analyze the Mahalanobis distance values.

```
psych::describe(item_scores_df$Mahal)
```

vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	66	3.81	2.24	3.68	3.62	2.36	0.2	11.25	11.05	0.86	0.82

Using this information we can determine cases that have a Mahalanobis distance values that exceeds three standard deviations around the median. In fact, we can have these noted in a column in the dataframe.

```

# creates a variable indicating TRUE or FALSE if an item is an
# outlier
item_scores_df$MOutlier <- dplyr::if_else(item_scores_df$Mahal > (median(item_scores_df$Mahal)
  (3 * sd(item_scores_df$Mahal))), TRUE, FALSE)

# shows us the first 6 rows of the data so we can see the new
# variables (Mahal, MOutlier)
head(item_scores_df)

# A tibble: 6 x 19
  ID iBIPOC_pr cmBlack Belong_1 Belong_2 Belong_3 Blst_1 Blst_2 Blst_3 Blst_4
  <int>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1     1      0.333      0       6       6       7       5       3       5       2
2     2      0.000      5       4       4       6       6       6       2       2
3     3      0.500     10      NA       3      NA      NA       5       2       2
4     4      0.333      6       5       3       2       2       2       2       2
5     5      1.000      5       4       4       4       6       1       1       1
6     6      0.000     20       5       6       5       5       1       1       2
# i 9 more variables: Blst_5 <dbl>, Blst_6 <dbl>, rBlst_1 <dbl>,
#   Belonging <dbl>, ResponseBL <dbl>, StigmaBL <dbl>, ClimateBL <dbl>,
#   Mahal <dbl>, MOutlier <lgl>

library(tidyverse)

-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.2      v readr     2.1.4
v forcats   1.0.0      v stringr   1.5.0
v ggplot2   3.4.3      v tibble    3.2.1
v lubridate 1.9.2      v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become

# counts frequency TRUE and FALSE indicating outlier or not
OutlierCount <- item_scores_df %>%
  dplyr::count(MOutlier)

# calculating how many outliers a slightly different way
nrow(item_scores_df) - OutlierCount

  MOutlier n
1       66  1
2       65  65

```

When we identify outliers we often ask if we should delete them or transform the data. A general rule of thumb is to look for “jumps” in the Mahalanobis distance values. If they are progressing steadily and there is no “jump,” researchers will often retain the outliers.

#### CUMULATIVE CAPTURE FOR THE APA STYLE WRITE-UP:

We evaluated multivariate normality with the Mahalanobis distance test. Specifically, we used the `psych::outlier()` function and included all continuous variables in the calculation. Our visual inspection of the Q-Q plot suggested that the plotted line strayed from the straight line as the quantiles increased. Additionally, we appended the Mahalanobis distance scores as a variable to the data. Analyzing this variable, we found that 1 case exceed three standard deviations beyond the median. Given that the Mahalanobis distance values increased in a consistent manner (i.e., no extreme “jumps”) we retained all cases.

## 3.7 A Few Words on Transformations

To quote from Kline [2016b], “Before applying a normalizing transformation, you should think about the variables of interest and whether the expectation of normality is reasonable.” (p. 77)

At this point in history, the non-normal distribution of the proportions of classmates who are Black and instructional staff who are BIPOC are accurate representations in higher education. Kline [2016b] has noted that transforming an inherently non-normal variable to force a normal distribution may fundamentally alter it such that the variable of interest is not actually studied. Kline’s chapter reviews some options for applying corrections to outliers. Additionally, the chapter describes a variety of normalizing transformations.

On a personal note, while I will use standardized scores (a linear transformation) if it improves interpretation and center variables around a meaningful intercept, I tend to resist the transformation of data without a really compelling reason. Why? It’s complicated and can make interpretation difficult.

## 3.8 The APA Style Write-Up

This results section will draw from the three lessons on scrubbing, scoring, and data diagnostics.:.

### 3.8.1 Data Diagnostics

Data screening suggested that 107 individuals opened the survey link. Of those, 83 granted consent and proceeded into the survey items. A further inclusion criteria was that the course was taught in the U.S; 69 met this criteria.

Available item analysis (AIA; [Parent, 2013]) is a strategy for managing missing data that uses available data for analysis and excludes cases with missing

data points only for analyses in which the data points would be directly involved. Parent (2013) suggested that AIA is equivalent to more complex methods (e.g., multiple imputation) across a number of variations of sample size, magnitude of associations among items, and degree of missingness. Thus, we utilized Parent’s recommendations to guide our approach to managing missing data. Missing data analyses were conducted with tools in base R as well as the R packages, *psych* (v. 2.3.6) and *mice* (v. 3.16.0).

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 67%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data. At this stage in the analysis, we allowed all cases with less than 90% missing to continue to the scoring stage. Guided by Parent’s [2013] AIA approach, scales with three items were scored if at least two items were non-missing; the scale with four items was scored if it at least three non-missing items; and the scale with six items was scored if it had at least five non-missing items.

Across the 66 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 61 cases. In this dataset we had less than 1% (0.55%) missing across the df; 97% of the rows had nonmissing data.

Regarding the distributional characteristics of the data, skew and kurtosis values of the variables fell below the values of 3 (skew) and 10 (kurtosis) that Kline suggests are concerning [2016b]. Results of the Shapiro-Wilk test of normality indicate that our variables assessing the proportion of classmates who are Black ( $W = 0.878, p < 0.001$ ) and the proportion of BIPOC instructional staff( $W = 0.787, p < 0.001$ ) are statistically significantly different than a normal distribution. The scales assessing the respondent’s belonging (0.973,  $p = 0.165$ ) and the respondent’s perception of campus climate for Black students ( $W = 0.951, p = 0.016$ ) did not differ differently from a normal distribution.

We evaluated multivariate normality with the Mahalanobis distance test. Specifically, we used the *psych::outlier()* function and included all continuous variables in the calculation. Our visual inspection of the Q-Q plot suggested that the plotted line strayed from the straight line as the quantiles increased. Additionally, we appended the Mahalanobis distance scores as a variable to the data. Analyzing this variable, we found that 1 case exceed three standard deviations beyond the median. Given that the Mahalanobis distance values increased in a consistent manner (i.e., no extreme “jumps”) we retained all cases.

Given that our sample sizes were reasonable for the planned analyses and the degree of missingness was low, we used pairwise deletion in our multiple regression analysis.

### 3.9 A Quick Regression of our Research Vignette

With some confidence that our scrubbed-and-scored variables are appropriate for analysis, let me conduct the super quick regression that is our research vignette.

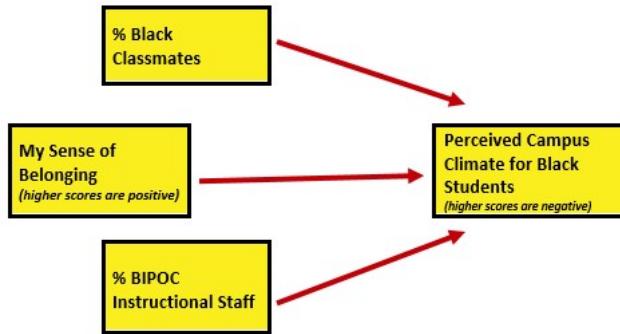


Figure 3.4: An image of the statistical model for which we are preparing data.

```
Climate_fit <- lm(ClimateBL ~ Belonging + cmBlack + iBIPOC_pr, data = item_scores_df)
summary(Climate_fit)
```

Call:

```
lm(formula = ClimateBL ~ Belonging + cmBlack + iBIPOC_pr, data = item_scores_df)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.86732	-0.80535	0.02355	0.70459	3.02003

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.90791	0.46653	6.233	0.0000000674 ***
Belonging	-0.01742	0.09643	-0.181	0.857
cmBlack	-0.01918	0.01717	-1.117	0.269
iBIPOC_pr	-0.64125	0.35701	-1.796	0.078 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.066 on 55 degrees of freedom

(7 observations deleted due to missingness)

Multiple R-squared: 0.08212, Adjusted R-squared: 0.03206

F-statistic: 1.64 on 3 and 55 DF, p-value: 0.1906

#### 3.9.1 Results

Results of a multiple regression predicting the respondents' perceptions of campus climate for Black students indicated that neither con-

butions of the respondents' personal belonging ( $B = -0.017, p = -.857$ ), the proportion of BIPOC instructional staff ( $B = 0.641, p = 0.078$ ), nor the proportion of Black classmates ( $B = -0.019, p = 0.269$ ) led to statistically significant changes in perceptions of campus climate for Black students. The model accounted for only 8% of the variance and was not statistically significant ( $p = 0.191$ ). Means, standard deviations, and correlations among variables are presented in Table 1; results of the regression model are presented in Table 2.

```
apaTables::apa.cor.table(item_scores_df[c("iBIPOC_pr", "cmBlack", "Belonging",
                                         "ClimateBL")], table.number = 1, show.sig.stars = TRUE, filename = "Table1_M_SDs_r_DataDx.rmd")
```

Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1	2	3
1. iBIPOC_pr	0.35	0.39			
2. cmBlack	8.20	8.02	.07 [-.18, .31]		
3. Belonging	4.03	1.47	.01 [-.24, .26]	-.13 [-.36, .12]	
4. ClimateBL	2.48	1.09	-.25 [-.47, .01]	-.17 [-.41, .08]	-.04 [-.29, .22]

Note. M and SD are used to represent mean and standard deviation, respectively.

Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

\* indicates  $p < .05$ . \*\* indicates  $p < .01$ .

```
library(apaTables)
apaTables::apa.reg.table(Climate_fit, table.number = 2, filename = "Climate_table.doc")
```

Table 2

Regression results using ClimateBL as the criterion

Predictor	b	b_95%_CI	beta	beta_95%_CI	sr2	sr2_95%_CI	r
(Intercept)	2.91**	[1.97, 3.84]					
Belonging	-0.02	[-0.21, 0.18]	-0.02	[-0.28, 0.24]	.00	[-.01, .01]	-.00
cmBlack	-0.02	[-0.05, 0.02]	-0.15	[-0.41, 0.12]	.02	[-.05, .09]	-.17
iBIPOC_pr	-0.64	[-1.36, 0.07]	-0.23	[-0.49, 0.03]	.05	[-.06, .16]	-.25

### Fit

R2 = .082  
95% CI [.00, .20]

Note. A significant b-weight indicates the beta-weight and semi-partial correlation are also significant. b represents unstandardized regression weights. beta indicates the standardized regression weight. sr2 represents the semi-partial correlation squared. r represents the zero-order correlation. Square brackets are used to enclose the lower and upper limits of a confidence interval. \* indicates p < .05. \*\* indicates p < .01.

## 3.10 Practice Problems

The three problems described below are designed to be continuations from the Scrubbing and Scoring lessons. You will likely encounter challenges that were not covered in this chapter. Search for and try out solutions, knowing that there are multiple paths through the analysis. The overall notion of the suggestions for practice are to (a) calculate alpha coefficients for the scales, (b) evaluate univariate and multivariate normality, (c) create an APA-style write-up appropriate for a data diagnostics subsection of the results, and (d) run a “quickie” regression, ANOVA, or similar analysis.

### 3.10.1 Problem #1: Reworking the Chapter Problem

If you chose this option in the prior chapters, you imported the data from Qualtrics, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, properly formatted the variables, interpreted item-level missingness, scored the scales/subscales, interpreted scale-level missingness, and wrote up the results. Please continue with the remaining tasks.

### 3.10.2 Problem #2: Use the *Rate-a-Recent-Course* Survey, Choosing Different Variables

If you chose this option in the prior chapter, you chose a minimum of three variables (different from those in the chapter) from the *Rate-a-Recent-Course* survey to include in a simple statistical model. You imported the data from Qualtrics, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, properly formatted the variables, interpreted item-level missingness, scored the scales/subscales, interpreted scale-level missingness, and wrote up the results. Please continue with the remaining tasks.

### 3.10.3 Problem #3: Other data

If you chose this option in the prior chapter, you used raw data that was available to you. You imported it into R, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, properly formatted the variables, interpreted item-level missingness, scored the scales/subscales, interpreted scale-level missingness, and wrote up the results. Please continue with the remaining tasks.

### 3.10.4 Grading Rubric

Assignment Component		
1. Calculate alpha coefficients for scales/subscales.	5	_____
2. Evaluate univariate normality (skew, kurtosis, Shapiro-Wilks).	5	_____
3. Evaluate multivariate normality (Mahalanobis test)	5	_____
4. Represent your work in an APA-style write-up (added to the writeup in the previous chapter)	5	_____
5. Conduct a quick analysis (e.g., regression, ANOVA) including at least three predictor variables	5	_____
6. Explanation to grader	5	_____
<b>Totals</b>	<b>30</b>	_____

## 3.11 Homeworked Example

### Screencast Link

For more information about the data used in this homeworked example, please refer to the description and codebook located at the end of the [introductory lesson](#) in [ReCentering Psych Stats](#). An .rds file which holds the data is located in the [Worked Examples](#) folder at the GitHub site the hosts the OER. The file name is *ReC.rds*.

Although the lessons focused on preparing data for analyses were presented in smaller sections, this homeworked example combines the suggestions for practice from the [Scrubbing](#), [Scoring](#), and [Data Dx](#) lessons. My hope is that this cumulative presentation is a closer approximation of what researchers need for their research projects.

These lessons were created to prepare a set of data to analyze a specific research model. Consequently, the model should be known and described at the beginning.

### 3.11.1 Scrubbing

#### Specify a research model

A further requirement was that the model should include three predictor variables (continuously or categorically scaled) and one dependent (continuously scaled) variable.

I am hypothesizing that socially responsive pedagogy (my dependent variable) will increase as a function of:

- the transition from SPSS (0) to R(1),
- the transition from a pre-centered (0) to re-centered (1) curriculum, and
- higher evaluations of traditional pedagogy

Because this data is nested within the person (i.e., students can contribute up to three course evaluations over the ANOVA, multivariate, and psychometrics courses) proper analysis would require a statistic (e.g., multilevel modeling) that would address the dependency in the data. Therefore, I will include only those students who are taking the multivariate modeling class.

*If you wanted to use this example and dataset as a basis for a homework assignment, you could create a different subset of data. I worked the example for students taking the multivariate modeling class. You could choose ANOVA or psychometrics. You could also choose a different combinations of variables.*

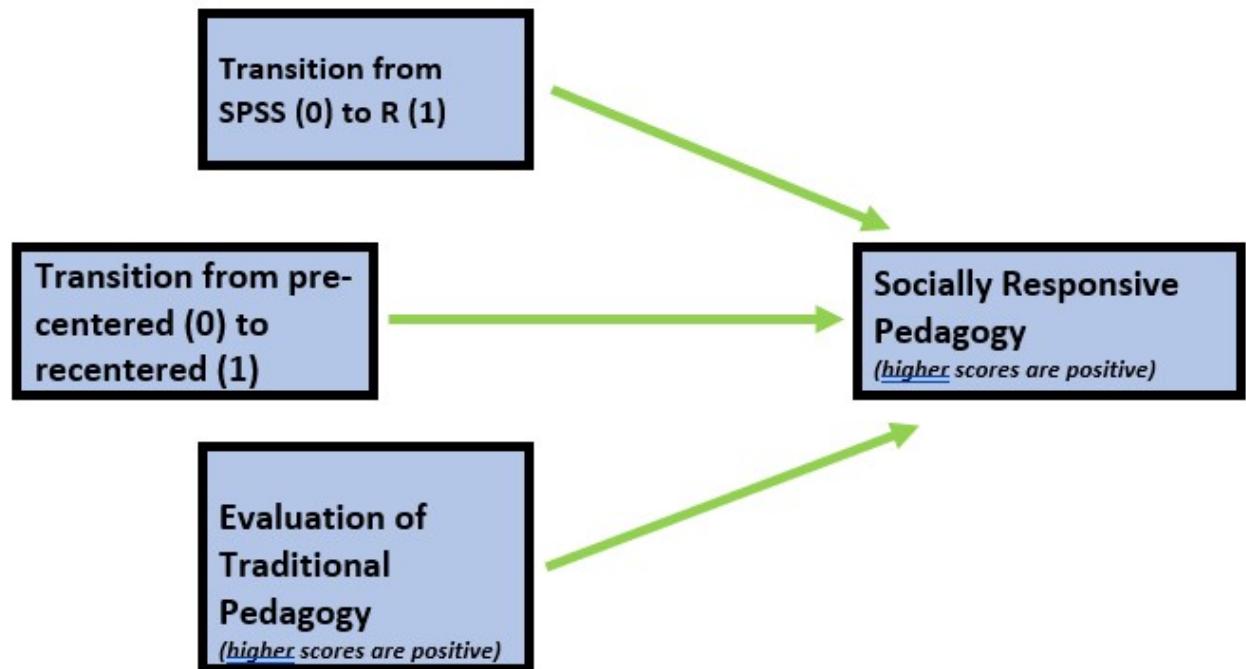


Figure 3.5: An image of our the prediction model for the homeworked example.

## Import data

```
raw <- readRDS("ReC.rds")
nrow(raw)
```

[1] 310

## Include only those who consented

Because this data is publicly posted on the Open Science Framework, it was necessary for me to already exclude those individuals. This data was unique in that students could freely write some version of “Opt out.” My original code included a handful of versions, but here was the basic form:

```
# testing to see if my code worked raw <- dplyr::filter (raw,
# SPFC.Decolonize.Opt.Out != 'Okay')
raw <- dplyr::filter(raw, SPFC.Decolonize.Opt.Out != "Opt Out")
```

## Apply exclusionary criteria

I want to exclude students’ responses for the ANOVA and psychometrics courses.

```
raw <- (dplyr::filter(raw, Course == "Multivariate"))
```

At this point, these my only inclusion/exclusion criteria. I can determine how many students (who consented) completed any portion of the survey.

```
nrow(raw)
```

[1] 84

## Rename variables to be sensible and systematic

Because this dataset is already on the OSF, the variables are sensibly named. However, I don’t like “SPFC.Decolonize.Opt.Out”. I will change it to simply “OptOut.”

```
raw <- dplyr::rename(raw, OptOut = "SPFC.Decolonize.Opt.Out")
```

It would have made more sense to do this before I used this variable in the calculations.

### Downsize the dataframe to the variables of interest

I will need to include:

- deID
- StatsPkg
- Centering
- Items included in the traditional pedagogy scale: ClearResponsibilities, EffectiveAnswers, Feedback, ClearOrganization, ClearPresentation
- Items included in the socially responsive pedagogy scale: InclusvClassrm, EquitableEval, MultPerspectives, DEIntegration

```
scrub_df <- (dplyr::select(raw, deID, StatsPkg, Centering, ClearResponsibilities,
                           EffectiveAnswers, Feedback, ClearOrganization, ClearPresentation, InclusvClassrm,
                           EquitableEval, MultPerspectives, DEIntegration))
```

### Provide an APA style write-up of these preliminary steps

This is a secondary analysis of data involved in a more comprehensive dataset that included students taking multiple statistics courses ( $N = 310$ ). Having retrieved this data from a repository in the Open Science Framework, only those who consented to participation in the study were included. Data used in these analyses were 84 students who completed the multivariate class.

#### 3.11.2 Scoring

##### Proper formatting of the item(s) in your first predictor variable

StatsPkg is a dichotomous variable. It should be structured as a factor with two ordered levels: SPSS, R

Because I am using the .rds form of the data from the OSF, this variable retains the former structure I assigned to it. If I needed to write the code, I would do this:

```
scrub_df$StatsPkg <- factor(scrub_df$StatsPkg, levels = c("SPSS", "R"))
str(scrub_df$StatsPkg)
```

Factor w/ 2 levels "SPSS","R": 2 2 2 2 2 2 2 2 2 ...

##### Proper formatting of item(s) in your second predictor variable

Similarly, Centering is a dichotomous variable. It should be structured as a factor with two ordered levels: Pre, Re.

Because I am using the .rds form of the data from the OSF, this variable retains the former structure I assigned to it. If I needed to write the code, I would do this:

```
scrub_df$Centering <- factor(scrub_df$Centering, levels = c("Pre", "Re"))
str(scrub_df$Centering)
```

Factor w/ 2 levels "Pre","Re": 2 2 2 2 2 2 2 2 2 2 ...

### Proper formatting of the item(s) in your third predictor variable

### Proper formatting of the item(s) in your dependent variable

The third predictor variable is traditional pedagogy. The dependent variable is socially responsive pedagogy. The items that will be used in the scale scores for both of these variables are all continuously scaled and should be identified as “int” or “num.” None of the items need to be reverse-scored.

```
str(scrub_df)
```

```
Classes 'data.table' and 'data.frame': 84 obs. of 12 variables:
 $ deID           : int 11 12 13 14 15 16 17 18 35 19 ...
 $ StatsPkg       : Factor w/ 2 levels "SPSS","R": 2 2 2 2 2 2 2 2 2 2 ...
 $ Centering      : Factor w/ 2 levels "Pre","Re": 2 2 2 2 2 2 2 2 2 2 ...
 $ ClearResponsibilities: int 4 5 5 5 4 3 5 5 3 5 ...
 $ EffectiveAnswers: int 4 5 5 4 4 3 5 5 4 4 ...
 $ Feedback        : int 4 5 4 4 5 4 5 4 4 5 ...
 $ ClearOrganization: int 3 5 5 4 4 3 5 5 4 5 ...
 $ ClearPresentation: int 4 5 5 3 4 2 5 4 5 5 ...
 $ InclusvClassrm   : int 5 5 5 5 5 4 5 5 5 5 ...
 $ EquitableEval    : int 4 5 5 5 4 4 5 4 5 5 ...
 $ MultPerspectives: int 4 5 5 5 5 5 5 4 5 5 ...
 $ DEIintegration   : int 5 5 5 5 5 5 5 5 5 5 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

### Evaluate and interpret item-level missingness

The *scrub\_df* is already downsized to include the item-level raw variables and the ID variable. We can continue using it.

I will create a “proportion missing” variable.

In this chunk I first calculate the number of missing (nmiss)

```
library(tidyverse)#needed because the script has pipes

#Calculating number and proportion of item-level missingness
scrub_df$nmiss <- scrub_df%>%
  dplyr::select(StatsPkg:DEIintegration) %>% #the colon allows us to include all variables b
  is.na %>%
```

```

rowSums

scrub_df<- scrub_df%>%
  dplyr::mutate(prop_miss = (nmiss/11)*100) #11 is the number of variables included in calculation

```

We can grab the descriptives for the `prop_miss` variable to begin to understand our data. I will create an object from it so I can use it with inline

```
psych::describe(scrub_df$prop_miss)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	84	2.38	6.17	0	0.94	0	0	36.36	36.36	3.29	12.33	0.67

Because I want to use the AIA approach to scoring, I'm not willing to filter out any cases yet. If I wanted to eliminate cases with egregious missing (i.e., like 90%), here is the code I would use:

```
scrub_df <- dplyr::filter(scrub_df, prop_miss <= 90) #update df to have only those with at least 10% missing
```

#### CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 36%.

To analyze missingness at the item level, we need a df that has only the variables of interest. That is, variables like `ID` and the `prop_miss` and `nmiss` variables we created will interfere with an accurate assessment of missingness. I will update our df to eliminate these.

```
# further update to exclude the n_miss and prop_miss variables
ItemMiss_df <- scrub_df %>%
  dplyr::select(-c(deID, nmiss, prop_miss))
```

Missing data analysis commonly looks at proportions by:

- the entire df
- rows/cases/people

```
# what proportion of cells missing across entire dataset
formattable::percent(mean(is.na(ItemMiss_df)))
```

[1] 2.38%

```
# what proportion of cases (rows) are complete (nonmissing)
formattable::percent(mean(complete.cases(ItemMiss_df)))
```

[1] 82.14%

#### CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 36%. Across the dataset, 2.38% of cells had missing data and 82.14% of cases had nonmissing data.

We can further explore patterns of missingness with *mice.md.pattern*.

```
mice:::md.pattern(ItemMiss_df, plot = TRUE, rotate.names = TRUE)
```

There are 6 missingness patterns. The most common ( $n = 69$ ) have no missingness. There are 11 students missing the DEIntegration item (on the traditional pedagogy scale). This item may have been a later addition to the Canvas course evaluations.

Comparing this to Enders' [2010] [prototypical patterns of missingness](#) (page 3), the *mice* output represents the monotonic pattern often caused by test fatigue. That is, once a student stopped responding, they didn't continue with the rest of the evaluation. That said, this was true of only 4 students (1 each pattern). A quick reminder – diagnosing monotonicity requires that the variables in the *mice.mdpattern* figures were presented to the research participant in that order.

#### Score any scales/subscales

Traditional pedagogy is a predictor variable that needs to be created by calculating the mean if at least 75% of the items are non-missing. None of the items need to be reverse-scored. I will return to working with the *scrub\_df* data.

```
# this seems to work when I build the book, but not in 'working the
# problem' TradPed_vars <- c('ClearResponsibilities',
# 'EffectiveAnswers', 'Feedback',
# 'ClearOrganization', 'ClearPresentation') scrub_df$TradPed <-
# sjstats::mean_n(scrub_df[, TradPed_vars], .75)

# this seems to work when I 'work the problem' (but not when I build
# the book) the difference is the two dots before the last SRPed_vars
TradPed_vars <- c("ClearResponsibilities", "EffectiveAnswers", "Feedback",
"ClearOrganization", "ClearPresentation")
scrub_df$TradPed <- sjstats::mean_n(scrub_df[, TradPed_vars], 0.75)
```

The dependent variable is socially responsive pedagogy. It needs to be created by calculating the mean if at least 75% of the items are non-missing. None of the items need to be reverse-scored.

```
# this seems to work when I build the book, but not in 'working the
# problem' SRPed_vars <- c('InclusvClassrm', 'EquitableEval',
# 'MultPerspectives', 'DEIintegration') scrub_df$SRPed <-
# sjstats::mean_n(scrub_df[, SRPed_vars], .75)

# this seems to work when I 'work the problem' (but not when I build
# the book) the difference is the two dots before the last SRPed_vars
SRPed_vars <- c("InclusvClassrm", "EquitableEval", "MultPerspectives",
  "DEIintegration")
scrub_df$SRPed <- sjstats::mean_n(scrub_df[, SRPed_vars], 0.75)
```

### Evaluate and interpret scale-level missingness

To evaluate scale level missingness, let's create a df with the focal variables.

```
scored <- dplyr::select(scrub_df, StatsPkg, Centering, TradPed, SRPed)
ScoredCaseMiss <- nrow(scored) #I produced this object for the sole purpose of feeding the nu
ScoredCaseMiss
```

[1] 84

Before we start our formal analysis of missingness at the scale level, let's continue to scrub by eliminating cases that will have too much missingness. In the script below we create a variable that counts the number of missing variables and then creates a proportion by dividing it by the number of total variables.

Using the *describe()* function from the *psych* package, we can investigate this variable.

```
library(tidyverse)
# Create a variable (n_miss) that counts the number missing
scored$n_miss <- scored %>%
  is.na %>%
  rowSums

# Create a proportion missing by dividing n_miss by the total number
# of variables (6) Pipe to sort in order of descending frequency to
# get a sense of the missingness
scored <- scored %>%
  mutate(prop_miss = (n_miss/6) * 100) %>%
  arrange(desc(n_miss))

psych::describe(scored$prop_miss)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	84	0.79	4.41	0	0	0	0	33.33	33.33	5.89	36.31	0.48

### CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 36%. Across the dataset, 2.38% of cells had missing data and 82.14% of cases had nonmissing data.

Across the 84 cases for which the scoring protocol was applied, missingness ranged from 0 to 33%.

We need to decide what is our retention threshold. Twenty percent seems to be a general rule of thumb. Let's delete all cases with missingness at 20% or greater.

```
# update df to have only those with at least 20% of complete data
# (this is an arbitrary decision)
scored <- dplyr::filter(scored, prop_miss <= 20)

# the variable selection just lops off the proportion missing
scored <- (select(scored, StatsPkg:SRPed))

# this produces the number of cases retained
nrow(scored)
```

[1] 83

### CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

Across the 84 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 83 cases.

Now, at the scale level, we look at missingness as the proportion of

- individual cells across the scored dataset, and
- rows/cases with nonmissing data

```
# percent missing across df
formattable::percent(mean(is.na(scored)))
```

[1] 0.60%

```
# percent of rows with nonmissing data  
formattable::percent(mean(complete.cases(scored)))
```

```
[1] 97.59%
```

#### CUMULATIVE CAPTURE FOR WRITING IT UP:

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

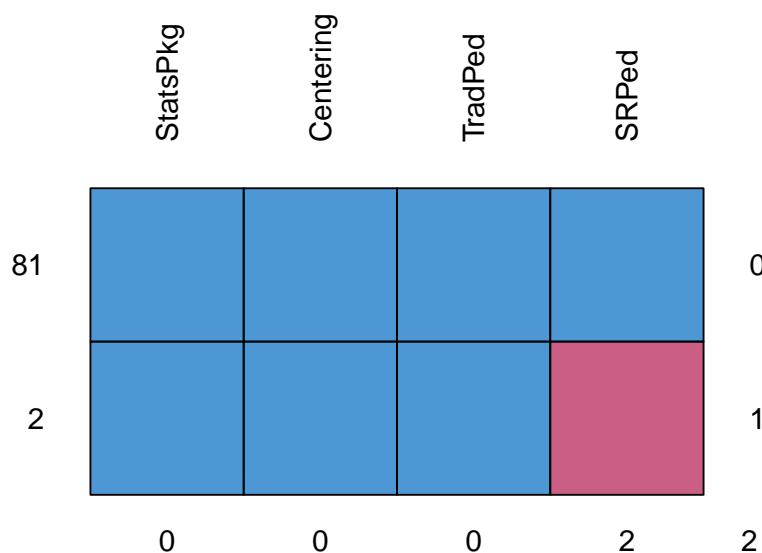
Across the 84 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 83 cases. In this dataset we had less than 1% (0.60%) missing across the df; 98% of the rows had nonmissing data.

Let's look again at missing patterns and mechanisms.

Returning to the *mice* package, we can use the *md.pattern()* function to examine a matrix with the number of columns 1 in which each row corresponds to a missing data pattern (0 = observed, 0 = missing). The rows and columns are sorted in increasing amounts of missing information. The last column and row contain row and column counts, respectively.

The corresponding figure shows non-missing data in blue; missing data in red.

```
mice_ScaleLvl <- mice::md.pattern(scored, plot = TRUE, rotate.names = TRUE)
```



```
mice_ScaleLvl
```

	StatsPkg	Centering	TradPed	SRPed
81	1	1	1	1 0
2	1	1	1	0 1
	0	0	0	2 2

There are 2 rows of data because there are only 2 patterns of missingness. The most common pattern is non-missing data ( $n = 81$ ). Two cases are missing the SRPed variable. If our statistical choice uses listwise deletion (i.e., the case is eliminated if one or more variables in the model has missing data), our sample size will be 79. As we will learn in later chapters, there are alternatives (i.e., specifying a FIML option in analyses that use maximum likelihood estimators) that can use all of the cases – even those with missing data.

**Represent your work in an APA-style write-up (added to the writeup in the previous chapter)**

Available item analysis (AIA; [Parent, 2013]) is a strategy for managing missing data that uses available data for analysis and excludes cases with missing data points only for analyses in which the data points would be directly involved. Parent (2013) suggested that AIA is equivalent to more complex methods (e.g., multiple imputation) across a number of variations of sample size, magnitude of associations among items, and degree of missingness. Thus, we utilized Parent's recommendations to guide our approach to managing missing data. Missing data analyses were conducted with tools in base R as well as the R packages, *psych* (v. 2.3.6) and *mice* (v. 3.16.0).

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

Across the 84 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 83 cases. In this dataset we had less than 1% (0.60%) missing across the df; 98% of the rows had nonmissing data.

### 3.11.3 Data Dx

#### Calculate alpha coefficients for scales/subscales

To calculate the alpha coefficients, we need item-level data. We will return to *scrub\_df* that contains the item-level data.

```
# alpha for the traditional pedagogy scale
psych::alpha(scrub_df[c("ClearResponsibilities", "EffectiveAnswers", "Feedback",
  "ClearOrganization", "ClearPresentation")])
```

```
Reliability analysis
Call: psych::alpha(x = scrub_df[c("ClearResponsibilities", "EffectiveAnswers",
  "Feedback", "ClearOrganization", "ClearPresentation")])

raw_alpha std.alpha G6(smc) average_r S/N    ase mean      sd median_r
  0.87        0.88     0.87      0.59 7.2 0.022   4.3 0.72      0.58

 95% confidence boundaries
      lower alpha upper
Feldt     0.83  0.87  0.91
Duhachek 0.83  0.87  0.92
```

Reliability if an item is dropped:

	raw_alpha	std.alpha	G6(smc)	average_r	S/N	alpha	se	var.r
ClearResponsibilities	0.84	0.84	0.82	0.57	5.3	0.029	0.0110	
EffectiveAnswers	0.84	0.84	0.81	0.57	5.2	0.029	0.0088	
Feedback	0.87	0.87	0.86	0.64	7.0	0.023	0.0053	
ClearOrganization	0.86	0.86	0.83	0.60	6.1	0.025	0.0067	
ClearPresentation	0.83	0.84	0.81	0.57	5.3	0.030	0.0074	
	med.r							
ClearResponsibilities	0.55							
EffectiveAnswers	0.58							
Feedback	0.63							
ClearOrganization	0.59							
ClearPresentation	0.57							

Item statistics

	n	raw.r	std.r	r.cor	r.drop	mean	sd
ClearResponsibilities	83	0.85	0.85	0.80	0.74	4.5	0.87
EffectiveAnswers	84	0.84	0.85	0.82	0.76	4.4	0.79
Feedback	82	0.74	0.75	0.65	0.60	4.3	0.81
ClearOrganization	84	0.82	0.80	0.74	0.68	4.1	1.04
ClearPresentation	84	0.85	0.85	0.81	0.76	4.2	0.87

Non missing response frequency for each item

	1	2	3	4	5	miss
ClearResponsibilities	0.01	0.05	0.04	0.27	0.64	0.01
EffectiveAnswers	0.02	0.00	0.05	0.40	0.52	0.00
Feedback	0.01	0.01	0.11	0.38	0.49	0.02
ClearOrganization	0.04	0.07	0.07	0.43	0.39	0.00
ClearPresentation	0.01	0.06	0.04	0.46	0.43	0.00

Cronbach's alpha for the traditional pedagogy scale was 0.88.

```
# alpha for the traditional pedagogy scale
psych::alpha(scrub_df[c("InclusvClassrm", "EquitableEval", "DEIintegration",
  "DEIIntegration")])
```

Warning in cor.smooth(r): Matrix was not positive definite, smoothing was done

```
In smc, smcs < 0 were set to .0
In smc, smcs < 0 were set to .0
In smc, smcs < 0 were set to .0
In smc, smcs < 0 were set to .0
```

Reliability analysis

```
Call: psych::alpha(x = scrub_df[c("InclusvClassrm", "EquitableEval",
```

```

"DEIintegration", "DEIintegration")])

raw_alpha std.alpha G6(smc) average_r S/N    ase mean    sd median_r
      0.85      0.85      0.7      0.58 5.6 0.025  4.5 0.62      0.55

95% confidence boundaries
      lower alpha upper
Feldt     0.79  0.85   0.9
Duhachek  0.80  0.85   0.9

Reliability if an item is dropped:
      raw_alpha std.alpha G6(smc) average_r S/N alpha se  var.r
InclusvClassrm      0.84      0.83      0.58      0.61 4.8  0.027 0.1115
EquitableEval       0.88      0.88      0.63      0.71 7.3  0.025 0.0640
DEIintegration      0.74      0.75      0.68      0.50 3.1  0.046 0.0054
DEIintegration.1    0.74      0.75      0.68      0.50 3.1  0.046 0.0054
      med.r
InclusvClassrm     0.42
EquitableEval      0.56
DEIintegration     0.53
DEIintegration.1   0.53

Item statistics
      n raw.r std.r r.cor r.drop mean    sd
InclusvClassrm    80  0.85  0.80  0.75  0.62  4.6  0.72
EquitableEval     84  0.71  0.72  0.60  0.51  4.7  0.50
DEIintegration    70  0.96  0.90  0.71  0.85  4.5  0.79
DEIintegration.1  70  0.96  0.90  0.71  0.85  4.5  0.79

Non missing response frequency for each item
      1    3    4    5 miss
InclusvClassrm   0.01 0.06 0.21 0.71 0.05
EquitableEval    0.00 0.01 0.32 0.67 0.00
DEIintegration   0.00 0.19 0.17 0.64 0.17
DEIintegration.1 0.00 0.19 0.17 0.64 0.17

```

Cronbach's alpha for the socially responsive pedagogy scale was 0.85.

Both of these are above the recommended value of 0.80.

### Evaluate univariate normality (skew, kurtosis, Shapiro-Wilks)

We can inspect univariate normality by examining the skew and kurtosis values of the continuously scored variables.

```
psych::describe(scored, type = 1)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
StatsPkg*	1	83	1.73	0.44	2.00	1.79	0.00	1.00	2	1.00	-1.06	-0.87
Centering*	2	83	1.36	0.48	1.00	1.33	0.00	1.00	2	1.00	0.58	-1.67
TradPed	3	83	4.29	0.72	4.40	4.40	0.59	1.20	5	3.80	-1.75	4.49
SRPed	4	81	4.51	0.58	4.75	4.60	0.37	2.33	5	2.67	-1.19	1.30
			se									
StatsPkg*			0.05									
Centering*			0.05									
TradPed			0.08									
SRPed			0.06									

When we use the “type=1” argument, the skew and kurtosis indices in the *psych* package can be interpreted according to Kline’s [2016a] guidelines.

Regarding the distributional characteristics of the data, skew and kurtosis values for our continuously scaled variables fall below the thresholds of concern (i.e., absolute value of 3 for skew; absolute value of 10 for kurtosis) identified by Kline [2016a].

Still at the univariate level, we can apply the Shapiro-Wilk test of normality to each of our continuously scaled variables. When the *p* value is < .05, the variable’s distribution is deviates from a normal distribution to a degree that is statistically significant. Below, the plotting of the histogram with a normal curve superimposed shows how the distribution approximates one that is normal.

```
# The shapiro-test is in base R; it's specification is simple:
# shapiro.test(df$variable) I added the object (and had to list it
# below) so I can use the inline text function
shapiro.test(scored$TradPed)
```

```
Shapiro-Wilk normality test

data: scored$TradPed
W = 0.83046, p-value = 0.0000000245

shapiro.test(scored$SRPed)
```

```
Shapiro-Wilk normality test

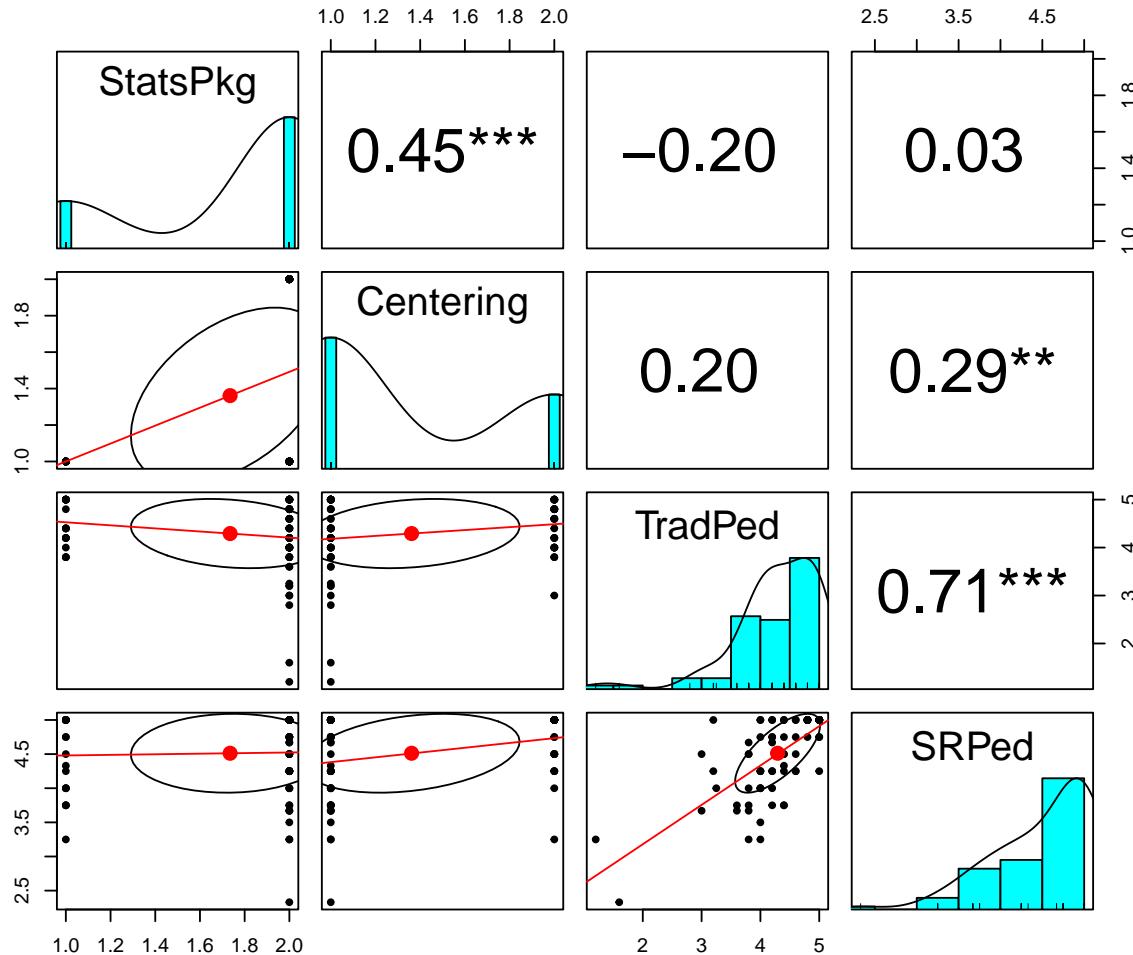
data: scored$SRPed
W = 0.81782, p-value = 0.0000000134
```

Both variable differ from a normal distribution in a statistically significant way.

- For the traditional pedagogy variable,  $W = 0.830, p < 0.001$
- for the socially responsive pedagogy variable,  $0.818, p < 0.001$

Obtaining a quick `psych::pairs.panel` can provide a quick glimpse of the distribution.

```
psych::pairs.panels(scored, stars = TRUE, lm = TRUE)
```



#### CUMULATIVE CAPTURE FOR THE APA STYLE WRITE-UP:

Regarding the distributional characteristics of the data, skew and kurtosis values of the variables fell below the values of 3 (skew) and 10 (kurtosis) that Kline suggests are concerning [2016b]. Results of the Shapiro-Wilk test of normality indicate that our variables assessing the traditional pedagogy ( $W = 0.830, p < 0.001$ ) and socially responsive pedagogy ( $0.818, p < 0.001$ ) are statistically significantly different than a normal distribution. Inspection

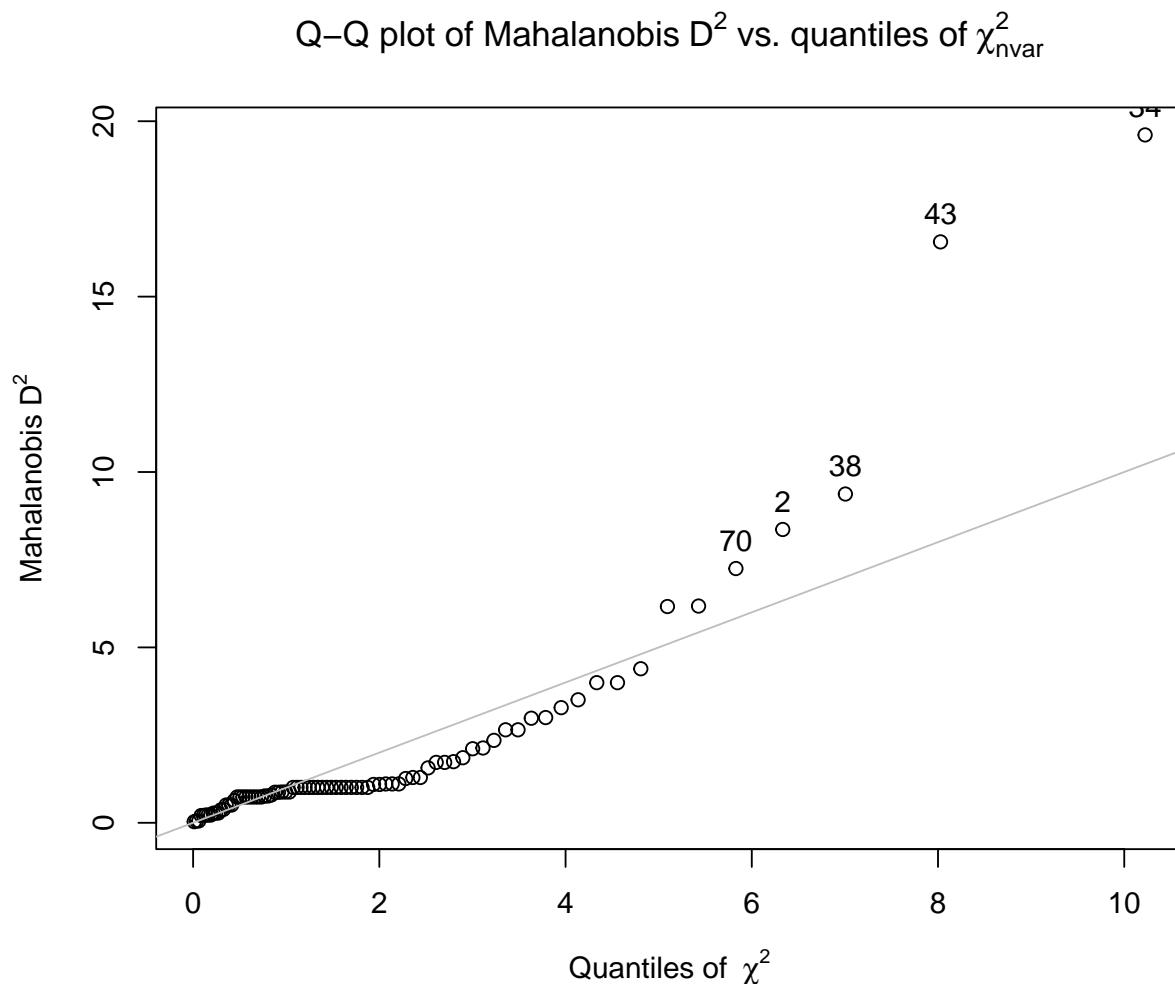
of distributions of the variables indicated that both course evaluation variables were negatively skewed, with a large proportion of high scores.

### Evaluate multivariate normality (Mahalanobis test)

In more complex models, multivariate normality is probably a more useful analysis. Although I am teaching this evaluation in advance of the formal analysis, as demonstrated in many of [ReCentering Psych Stats ANOVA chapters](#), this can also be assessed by examining the distribution of residuals after the analysis is complete.

Multivariate normality can be assessed with the continuously scaled variables. The code below includes the only two continuously scaled variables. The code simultaneously (a) appends the df with a Mahalanobis value and (b) creates a QQ plot. Dots that stray from the line are the scores that are contributing to multivariate non-normality.

```
scored$Mahal <- psych::outlier(scored[c("TradPed", "SRPed")])
```



We can analyze the distributional characteristics of the Mahalanobis values with *psych::describe*. It is possible, then to analyze the Mahalanobis distance values.

```
psych::describe(scored$Mahal)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	
X1		1	83	1.97	3.12	1.01	1.27	0.42	0.03	19.61	19.58	3.75	15.87	0.34

Using this information we can determine cases that have a Mahalanobis distance values that exceeds three standard deviations around the median. In fact, we can have these noted in a column in the dataframe.

```
# creates a variable indicating TRUE or FALSE if an item is an
# outlier
scored$MOutlier <- dplyr::if_else(scored$Mahal > (median(scored$Mahal) +
  (3 * sd(scored$Mahal))), TRUE, FALSE)
```

```
# shows us the first 6 rows of the data so we can see the new
# variables (Mahal, MOutlier)
head(scored)
```

	StatsPkg	Centering	TradPed	SRPed	Mahal	MOutlier
1	SPSS	Pre	4.2	NA	0.0319020	FALSE
2	R	Pre	2.8	NA	8.3615550	FALSE
3	R	Re	3.8	4.5	0.8702516	FALSE
4	R	Re	5.0	5.0	1.0087776	FALSE
5	R	Re	4.8	5.0	0.7363631	FALSE
6	R	Re	4.0	5.0	2.6509906	FALSE

```
library(tidyverse)
# counts frequency TRUE and FALSE indicating outlier or not
OutlierCount <- scored %>%
  dplyr::count(MOutlier)

# calculating how many outliers a slightly different way
nrow(scored) - OutlierCount
```

	MOutlier	n
1	83	2
2	82	81

When we identify outliers we often ask if we should delete them or transform the data. A general rule of thumb is to look for “jumps” in the Mahalanobis distance values. If they are progressing steadily and there is no “jump,” researchers will often retain the outliers.

In this case, I do see a jump. When I sort the df on Mahal values, the jump from 9.37 to 16.56 is much different than the more gradual increase in values that precedes it. Therefore, I think I will delete cases with Mahalanobis values greater than 10 (a number I “just picked”).

```
scored <- dplyr::filter(scored, Mahal < "10")
```

We evaluated multivariate normality with the Mahalanobis distance test. Specifically, we used the *psych::outlier()* function and included both continuous variables in the calculation. Our visual inspection of the Q-Q plot suggested that the plotted line strayed from the straight line as the quantiles increased. Additionally, we appended the Mahalanobis distance scores as a variable to the data. Analyzing this variable, we found that 2 cases exceed three standard deviations beyond the median. Because there was a substantial “jump” between the non-outliers and these two variables we chose to delete them.

### **Represent your work in an APA-style write-up (added to the writeup in the previous chapter)**

This is a secondary analysis of data involved in a more comprehensive dataset that included students taking multiple statistics courses ( $N = 310$ ). Having retrieved this data from a repository in the Open Science Framework, only those who consented to participation in the study were included. Data used in these analyses were 84 students who completed the multivariate clas.

Available item analysis (AIA; [Parent, 2013]) is a strategy for managing missing data that uses available data for analysis and excludes cases with missing data points only for analyses in which the data points would be directly involved. Parent (2013) suggested that AIA is equivalent to more complex methods (e.g., multiple imputation) across a number of variations of sample size, magnitude of associations among items, and degree of missingness. Thus, we utilized Parent’s recommendations to guide our approach to managing missing data. Missing data analyses were conducted with tools in base R as well as the R packages, *psych* (v. 2.3.6) and *mice* (v. 3.16.0).

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data.

Across the 84 cases for which the scoring protocol was applied, missingness ranged from 0 to 67%. After eliminating cases with greater than 20% missing, the dataset analyzed included 83 cases. In this dataset we had less than 1% (0.60%) missing across the df; 98% of the rows had nonmissing data.

Regarding the distributional characteristics of the data, skew and kurtosis values of the variables fell below the values of 3 (skew) and 10 (kurtosis) that Kline suggests are concerning [2016b]. Results of the Shapiro-Wilk test of normality indicate that our variables assessing the traditional pedagogy ( $W = 0.830, p < 0.001$ ) and socially responsive pedagogy ( $0.818, p < 0.001$ )

are statistically significantly different than a normal distribution. Inspection of distributions of the variables indicated that both course evaluation variables were negatively skewed, with a large proportion of high scores.

We evaluated multivariate normality with the Mahalanobis distance test. Specifically, we used the *psych::outlier()* function and included both continuous variables in the calculation. Our visual inspection of the Q-Q plot suggested that the plotted line strayed from the straight line as the quantiles increased. Additionally, we appended the Mahalanobis distance scores as a variable to the data. Analyzing this variable, we found that 2 cases exceed three standard deviations beyond the median. Because there was a substantial “jump” between the non-outliers and these two variables we chose to delete them.

**Conduct a quick analysis (e.g., regression, ANOVA) including at least three variables**

```
SRPed_fit <- lm(SRPed ~ StatsPkg + Centering + TradPed, data = scored)
summary(SRPed_fit)
```

Call:

```
lm(formula = SRPed ~ StatsPkg + Centering + TradPed, data = scored)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.56099	-0.14406	0.01551	0.10594	0.46498

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.46330	0.34441	4.249	0.000077464849487 ***
StatsPkgR	0.13251	0.08056	1.645	0.105
CenteringRe	0.05666	0.07423	0.763	0.448
TradPed	0.68663	0.07365	9.323	0.000000000000332 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2433 on 59 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.6167, Adjusted R-squared: 0.5972

F-statistic: 31.64 on 3 and 59 DF, p-value: 0.000000000002547

### 3.11.4 Results

Results of a multiple regression predicting the socially responsive course evaluation ratings indicated that neither the transition from SPSS to R ( $B =$

$0.133, p = 0.105$ ) nor the transition to an explicitly recentered curriculum ( $B = 0.057, p = 0.448$ ) led to statistically significant differences. In contrast, traditional pedagogy had a  $\beta = 0.686$ ,  $p < 0.001$ ). The model accounted for 62% of the variance and was statistically significant ( $p, 0.001$ ). Means, standard deviations, and correlations among variables are presented in Table 1; results of the regression model are presented in Table 2.

```
apaTables::apa.cor.table(scored[c("SRPed", "StatsPkg", "Centering", "TradPed")],  
  table.number = 1, show.sig.stars = TRUE, filename = "Table1__DataDx_HW.doc")
```

Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1
1. SRPed	4.69	0.38	
2. TradPed	4.53	0.43	.76** [.63, .85]

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

\* indicates  $p < .05$ . \*\* indicates  $p < .01$ .

```
apaTables::apa.reg.table(SRPed_fit, table.number = 2, filename = "SRPed_table.doc")
```

Table 2

Regression results using SRPed as the criterion

Predictor	b	b_95%_CI	sr2	sr2_95%_CI	Fit
(Intercept)	1.46**	[0.77, 2.15]			
StatsPkgR	0.13	[-0.03, 0.29]	.02	[-.02, .06]	
CenteringRe	0.06	[-0.09, 0.21]	.00	[-.02, .02]	
TradPed	0.69**	[0.54, 0.83]	.56	[.40, .73]	R2 = .617**

95% CI [.43,.70]

Note. A significant b-weight indicates the semi-partial correlation is also significant.  
b represents unstandardized regression weights.

sr2 represents the semi-partial correlation squared.

Square brackets are used to enclose the lower and upper limits of a confidence interval.

\* indicates  $p < .05$ . \*\* indicates  $p < .01$ .



# Chapter 4

## Multiple Imputation (A Brief Demo)

### [Screencasted Lecture Link](#)

Multiple imputation is a tool for managing missing data that works with the whole raw data file to impute values for missing data for *multiple sets* (e.g., 5-20) of the raw data. Those multiple sets are considered together in analyses (such as regression) and interpretation is made on the pooled results. Much has been written about multiple imputation and, if used, should be done with many considerations. This chapter is intended as a brief introduction. In this chapter, I demonstrate the use of multiple imputation with the data from the [Rate-a-Recent-Course: A ReCentering Psych Stats Exercise](#) that has served as the research vignette for the first few chapters of this OER.

### 4.1 Navigating this Lesson

There is about one hour of lecture. If you work through the materials with me it would be good to add another hour (to an hour-and-a-half).

While the majority of R objects and data you will need are created within the R script that sources the chapter, there are a few that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

#### 4.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Describe circumstances under which multiple imputation would be appropriate
- List and define the stages in multiple imputation.
- Apply multiple imputation to a dataset that has missingness
- Interpret results from a simple regression that uses multiple imputation
- Articulate how multiple imputation fits into the workflow for scrubbing and scoring data.
- Write up the results of an the process of imputation from raw data through analyzing a simple regression (or similar) analysis.

### 4.1.2 Planning for Practice

The suggestions for practice are a continuation from the three prior chapters. If you have completed one or more of those assignments, you should have worked through the steps in preparing a data set and evaluating its appropriateness for the planned, statistical, analysis. This chapter takes a deviation from the AIA [Parent, 2013] approach that was the focus of the first few chapters in that we used multiple imputation as the approach for managing missingness. Options, of graded complexity, for practice include:

- Repeating the steps in the chapter with the most recent data from the Rate-A-Recent-Course survey; differences will be in the number of people who have completed the survey since the chapter was written.
- Use the dataset that is the source of the chapter, but score a different set of items that you choose.
- Begin with raw data to which you have access.

### 4.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98, 4–18.
  - Craig Enders is a leading expert in the analysis and management of missing data. This article is useful in describing multiple imputation as a method for managing missingness.
- Katitas, A. (2019). Getting Started with Multiple Imputation in R. University of Virginia Library: Research Data Services + Sciences. <https://library.virginia.edu/data/articles/getting-started-with-multiple-imputation-in-r>
  - Tutorial for conducting multiple imputation in R.
- Kline Ch4, Data Preparation & Psychometrics Review (pp. 72/Outliers - 88/Modern Methods)
- Kline's chapter is my “go-to” for making decisions about preparing data for analysis.

### 4.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
if (!require(qualtRics)) {
  install.packages("qualtRics")
}
if (!require(psych)) {
  install.packages("psych")
```

```

}
if (!require(dplyr)) {
  install.packages("dplyr")
}
if (!require(mice)) {
  install.packages("mice")
}

```

## 4.2 Workflow for Multiple Imputation

The following is a proposed workflow for preparing data for analysis.

In this lecture we are working on the right side of the flowchart in the multiple imputation (blue) section. Within it, there are two options, each with a slightly different set of options.

- imputing at the item level
  - in this case, scales/subscales are scored after the item-level imputation
- imputating at the scale level
  - in this case, scales/subscales are scored prior to the imputation; likely using some of the same criteria as identified in the scoring chapter (i.e., scoring if 75-80% of data are non-missing). Multiple imputation, then, is used to estimate the remaining, missing values.

Whichever approach is used, the imputed variables (multiple sets) are used in a *pooled analysis* and results are interpreted from that analysis.

## 4.3 Research Vignette

The research vignette comes from the survey titled, [Rate-a-Recent-Course: A ReCentering Psych Stats Exercise](#) and is explained in the [scrubbing chapter](#). In the [scoring chapter](#) we prepared four variables for analysis. In the [data diagnostics chapter](#) we assessed the quality of the variables and conducted the multiple regression described below. Details for these are in our [codebook](#).

Let's quickly review the variables in our model:

- Perceived Campus Climate for Black Students includes 6 items, one of which was reverse scored. This scale was adapted from Szymanski et al.'s [2020] Campus Climate for LGBTQ students. It has not been evaluated for use with other groups. The Szymanski et al. analysis suggested that it could be used as a total scale score, or divided into three items each that assess
  - College response to LGBTQ students (items 6, 4, 1)
  - LGBTQ stigma (items 3, 2, 5)

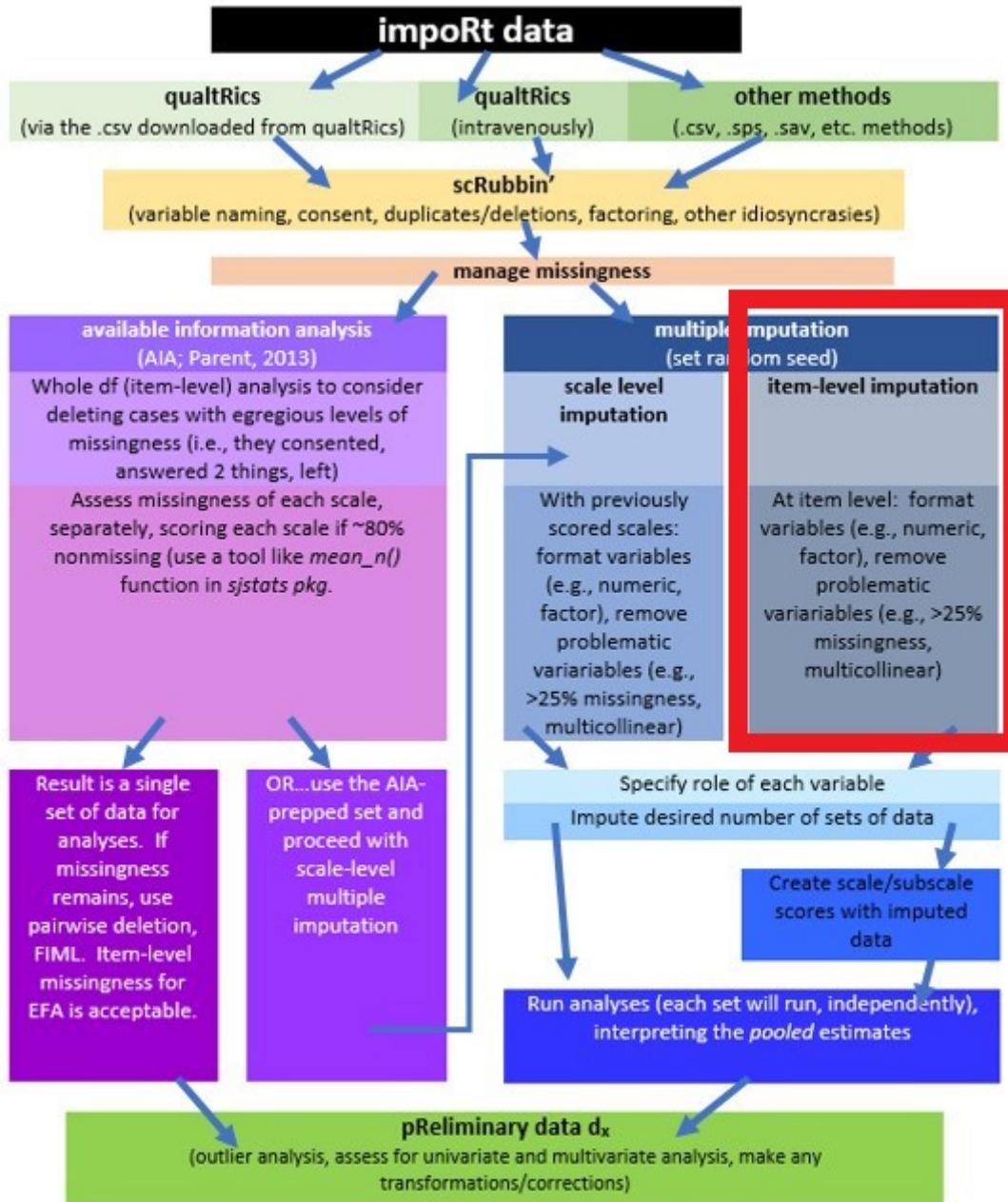


Figure 4.1: An image of a workflow for scrubbing and scoring data.

- Sense of Belonging includes 3 items. This is a subscale from Bollen and Hoyle's [1990] Perceived Cohesion Scale. There are no items on this scale that require reversing.
- Percent of Black classmates is a single item that asked respondents to estimate the proportion of students in various racial categories
- Percent of BIPOC instructional staff, similarly, asked respondents to identify the racial category of each member of their instructional staff

As we noted in the [scrubbing chapter](#), our design has notable limitations. Briefly, (a) owing to the open source aspect of the data we do not ask about the demographic characteristics of the respondent; (b) the items that ask respondents to *guess* the identities of the instructional staff and to place them in broad categories, (c) we do not provide a “write-in” a response. We made these decisions after extensive conversation with stakeholders. The primary reason for these decisions was to prevent potential harm (a) to respondents who could be identified if/when the revealed private information in this open-source survey, and (b) trolls who would write inappropriate or harmful comments.

As I think about “how these variables go together” (which is often where I start in planning a study), I suspect parallel mediation. That is the perception of campus climate for Black students would be predicted by the respondent’s sense of belonging, mediated in separate paths through the proportion of classmates who are Black and the proportion of BIPOC instructional staff.

*I would like to assess the model by having the instructional staff variable to be the %Black instructional staff. At the time that this lecture is being prepared, there is not sufficient Black representation in the instructional staff to model this.*

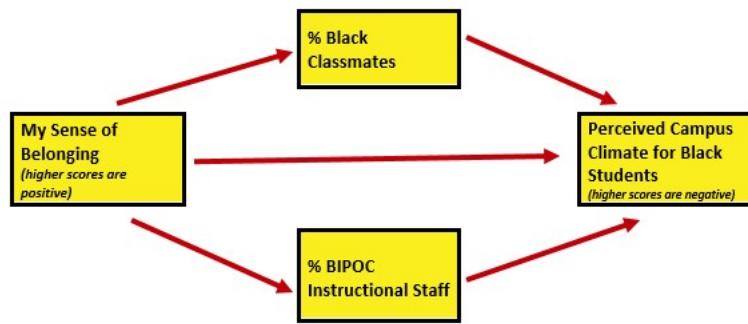


Figure 4.2: An image of the statistical model for which we are preparing data.

As in the [data diagnostic chapter](#), I will conclude this chapter by conducting a statistical analysis with the multiply imputed data. Because parallel mediation can be complicated (I teach it in a later chapter), I will demonstrate use of our prepared variables with a simple multiple regression.

## 4.4 Multiple Imputation – a Super Brief Review

Multiple imputation is complex. Numerous quantitative psychologists had critiqued it and provided numerous cautions and guidelines for its use [[Enders, 2010, 2017](#), [Little et al., 2008](#), [Little and Rubin, 2002](#)]. In brief,

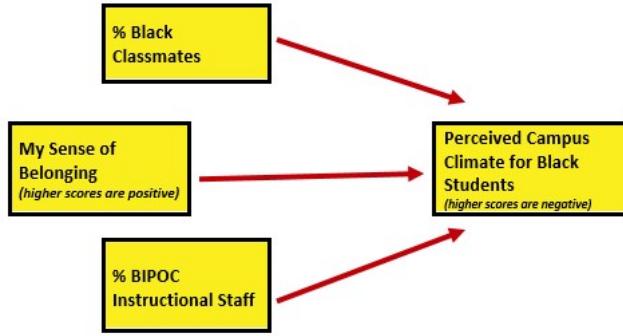


Figure 4.3: An image of the statistical model for which we are preparing data.

#### 4.4.1 Steps in Multiple Imputation

- Multiple imputation starts with a raw data file.
  - Multiple imputation assumes that data are MAR (remember, MCAR is the more prestigious one). This means that researchers assume that missing values can be replaced by predictions derived from the observable portion of the dataset.
- Multiple datasets (often 5 to 20) are created where missing values are replaced via a randomized process (so the same missing value [item 4 for person A] will likely have different values for each dataset).
- The desired analysis is conducted simultaneously/separately for each of the imputed sets (so if you imputed 5 sets and wanted a linear regression, you get 5 linear regressions).
- A *pooled analysis* uses the point estimates and the standard errors to provide a single result that represents the analysis.

In a web-hosted guide from the University of Virginia Library, Katitas [2019] provided a user-friendly review and example of using tools in R in a multiple imputation. Katitas' figure is a useful conceptual tool in understanding how multiple imputation works. *This figure is my recreation of Katitas' original.*

- the dataframe with missing data is the single place we start
- we intervene with a package like `mice()` to
- impute multiple sets of data (filling in the missing variables with different values that are a product of their conditional distribution and an element of “random”);
  - “mids” (“multiply imputed dataset”) is an object class where the completed datasets are stored.
- the “with\_mids” command allows OLS regression to be run, as many times as we have imputed datasets (in this figure, 3X). It produces different regression coefficients for each dataset

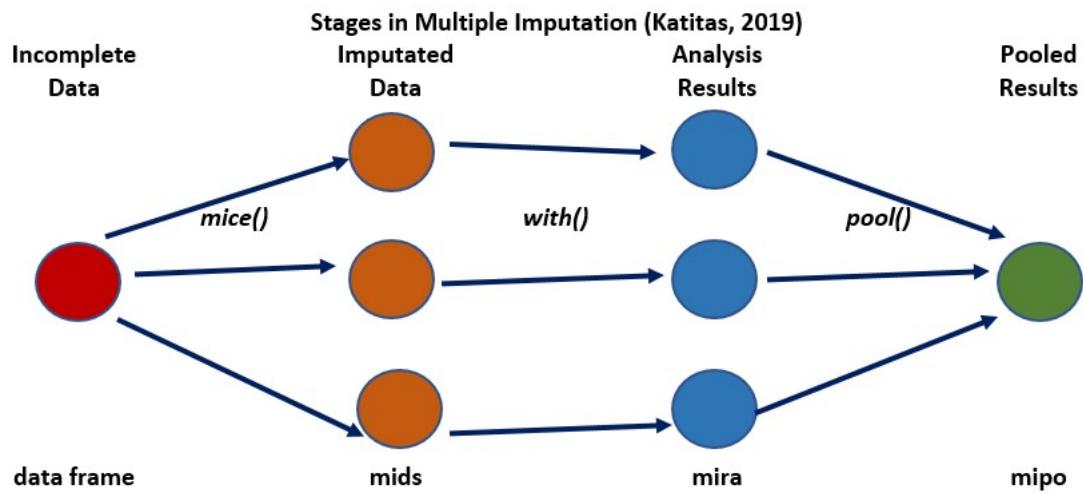


Figure 4.4: An image adapted from the Katitas multiple imputation guide showing the four stages of multiple imputation.

- the “pool” command pools together the multiple coefficients taking into consideration the value of the coefficients, the standard errors, and the variance of the missing value parameter across the samples.

#### 4.4.2 Statistical Approaches to Multiple Imputation

**Joint multivariate normal distribution multiple imputation** assumes that the observed data follow a multivariate normal distribution. The algorithm used draws from this assumed distribution. A drawback is that if the data do not follow a multivariate normal distribution, the imputed values are incorrect. *Amelia* and *norm* packages use this approach.

**Conditional multiple imputation** is an iterative procedure, modeling the conditional distribution of a certain variable given the other variables. In this way the distribution is assumed for each variable, rather than for the entire dataset. *mice* uses this approach.

*mice*: multivariate imputation by chained equations

## 4.5 Working the Problem

Katitas [2019] claims that it is best to impute the data in its rawest form possible because any change would be taking it away from its original distribution. There are debates about how many variables to include in an imputation. Some authors would suggest that researchers include everything that was collected. Others (like me) will trim the dataset to include (a) the variables included in the model, plus (b) auxiliary variables (i.e., variables not in the model, but that are sufficiently non-missing and will provide additional information to the data).

In our case we will want:

Item for the variables represented in our model

- the item level responses to the scales/subscales
  - respondents' sense of belonging to campus (3 items)
  - respondents' rating of campus climate for Black students (6 items)
- proportion of BIPOC instructional staff
- proportion of classmates who are Black

Auxiliary variables – let's choose four. One will be the format of the course. Three items will be from the course evaluation.

- format, whether the course was taught in-person, a blend, or virtual
- cEval\_1, "Course material was presented clearly"
- cEval\_13, "Where applicable, issues were considered from multiple perspectives"
- cEval\_19, "My understanding of the subject matter increased over the span of the course"

#### 4.5.1 Selecting and Formatting Variables

There are some guidelines for selecting and formatting variables for imputation.

- Variables should be in their *most natural* state
- Redundant or too highly correlated variables should not be included
  - If you reverse coded a variable (we haven't yet), that's ok, but if you have already reverse-coded, then exclude the original variable
  - Redundant variables (or multicollinear variables) may cause the multiple imputation process to cease
  - Violation of this also provides clues for troubleshooting
- Exclude variables with more than 25% missing

To make this as realistic as possible. Let's start with our very raw data. The [Scrubbing chapter](#) provides greater detail on importing data directly from Qualtrics. If you have worked the lessons, consecutively, you know that data can be added to this survey at any time. So that the values in the chapter are consistent, I will use the datafiles that I immediately saved when I conducted the analysis at the time I last updated the chapter.

Please download the .rds or .csv file from [MultivModel GitHub](#) site. Please the file in the same folder as your .rmd file. As always, I prefer working with .rds files.

```
QTRX_df2 <- readRDS("QTRX_df230902b.rds")
# QTRX_df <- read.csv('QTRX_df230902b.csv', header = TRUE)
```

Next, I apply inclusion/exclusion criteria. As described in the [Scrubbing chapter](#) this includes:

- excluding all *previews*

- including only those who consented
- including only those whose rated course was offered by a U.S. institution

```
library(tidyverse)
QTRX_df2 <- dplyr::filter(QTRX_df2, DistributionChannel != "preview")
QTRX_df2 <- dplyr::filter(QTRX_df2, Consent == 1)
QTRX_df2 <- dplyr::filter(QTRX_df2, USInst == 0)
```

Preparing the data also meant renaming some variables that started with numbers (a hassle in R). I also renamed variables on the Campus Climate scale so that we know to which subscale they belong.

```
# renaming variables that started with numbers
QTRX_df2 <- dplyr::rename(QTRX_df2, iRace1 = "1_iRace", iRace2 = "2_iRace",
                           iRace3 = "3_iRace", iRace4 = "4_iRace", iRace5 = "5_iRace", iRace6 = "6_iRace",
                           iRace7 = "7_iRace", iRace8 = "8_iRace", iRace9 = "9_iRace", iRace10 = "10_iRace")
# renaming variables from the identification of classmates
QTRX_df2 <- dplyr::rename(QTRX_df2, cmBiMulti = Race_10, cmBlack = Race_1,
                           cmNBPoC = Race_7, cmWhite = Race_8, cmUnsure = Race_2)
```

The Qualtrics download does not include an ID number. Because new variables are always appended to the end of the df, we also include code to make this the first column.

```
QTRX_df2 <- QTRX_df2 %>%
  dplyr::mutate(ID = row_number())
# moving the ID number to the first column; requires
QTRX_df2 <- QTRX_df2 %>%
  dplyr::select(ID, everything())
```

Because this huge df is cumbersome to work with, let's downsize it to be closer to the size we will work with in the imputation

```
mimp_df <- dplyr::select(QTRX_df2, ID, iRace1, iRace2, iRace3, iRace4,
                           iRace5, iRace6, iRace7, iRace8, iRace9, iRace10, cmBiMulti, cmBlack,
                           cmNBPoC, cmWhite, cmUnsure, Belong_1:Belong_3, Blst_1:Blst_6, cEval_1,
                           cEval_13, cEval_19, format)
# glimpse(mimp_df)
head(mimp_df)
```

```
# A tibble: 6 x 29
  ID iRace1 iRace2 iRace3 iRace4 iRace5 iRace6 iRace7 iRace8 iRace9 iRace10
  <int> <dbl> <dbl> <dbl> <dbl> <lgl> <lgl> <lgl> <lgl> <lgl> <lgl>
1     1      3      1      3     NA  NA     NA     NA     NA  NA     NA
2     2      3     NA     NA     NA  NA     NA     NA     NA  NA     NA
3     3      3      1     NA     NA  NA  NA     NA     NA  NA     NA
4     4      3      1      3     NA  NA     NA     NA     NA  NA     NA
```

```

5      5      1      NA      NA      NA NA      NA      NA      NA      NA
6      6      3      NA      NA      NA NA      NA      NA      NA      NA
# i 18 more variables: cmBiMulti <dbl>, cmBlack <dbl>, cmNBPoC <dbl>,
#   cmWhite <dbl>, cmUnsure <dbl>, Belong_1 <dbl>, Belong_2 <dbl>,
#   Belong_3 <dbl>, Blst_1 <dbl>, Blst_2 <dbl>, Blst_3 <dbl>, Blst_4 <dbl>,
#   Blst_5 <dbl>, Blst_6 <dbl>, cEval_1 <dbl>, cEval_13 <dbl>, cEval_19 <dbl>,
#   format <dbl>

```

#### 4.5.2 Creating Composite Variables

Qualtrics imports many of the categorical variables as numbers. R often reads them numerically (integers or numbers). If they are directly converted to factors, R will sometimes collapse. In this example, if there is a race that is not represented (e.g., 2 for BiMulti), when the numbers are changed to factors, R will assume it's ordered and will change up the numbers. Therefore, it is ESSENTIAL to check (again and again ad nauseum) to ensure that your variables are recoding in a manner you understand.

```

mimp_df$iRace1 = factor(mimp_df$iRace1, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace2 = factor(mimp_df$iRace2, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace3 = factor(mimp_df$iRace3, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace4 = factor(mimp_df$iRace4, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace5 = factor(mimp_df$iRace5, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace6 = factor(mimp_df$iRace6, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace7 = factor(mimp_df$iRace7, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace8 = factor(mimp_df$iRace8, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace9 = factor(mimp_df$iRace9, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))
mimp_df$iRace10 = factor(mimp_df$iRace10, levels = c(0, 1, 2, 3, 4), labels = c("Black",
  "nBpoc", "BiMulti", "White", "NotNotice"))

head(mimp_df)

```

This is a quick recap of how we calculated the proportion of instructional staff who are BIPOC.

```

# creating a count of BIPOC faculty identified by each respondent
mimp_df$count.BIPOC <- apply(mimp_df[c("iRace1", "iRace2", "iRace3", "iRace4",
  "iRace5", "iRace6", "iRace7", "iRace8", "iRace9", "iRace10")], 1, function(x) sum(x %in%
  c("Black", "nBpoc", "BiMulti")))

```

```
# creating a count of all instructional faculty identified by each
# respondent
mimp_df$count.nMiss <- apply(mimp_df[c("iRace1", "iRace2", "iRace3", "iRace4",
  "iRace5", "iRace6", "iRace7", "iRace8", "iRace9", "iRace10")], 1, function(x) sum(!is.na(x))

# calculating the proportion of BIPOC faculty with the counts above
mimp_df$iBIPOC_pr = mimp_df$count.BIPOC/mimp_df$count.nMiss
```

I have included another variable, *format* that we will use as auxiliary variable. As written, these are the following meanings:

1. In-person (all persons are attending in person)
2. In person (some students are attending remotely)
3. Blended: some sessions in person and some sessions online/virtual
4. Online or virtual
5. Other

Let's recode it to have three categories:

0. 100% in-person (1)
1. Some sort of blend/mix (2, 3)
2. 100% online/virtual (4) NA. Other (5)

```
# we can assign more than one value to the same factor by repeating
# the label
mimp_df$format = factor(mimp_df$format, levels = c(1, 2, 3, 4, 5), labels = c("InPerson",
  "Blend", "Blend", "Online", is.na(5)))
```

Let's trim the df again to just include the variables we need in the imputation.

```
mimp_df <- select(mimp_df, ID, iBIPOC_pr, cmBlack, Belong_1:Belong_3, Blst_1:Blst_6,
  cEval_1, cEval_13, cEval_19, format)
```

Recall one of the guidelines was to remove variables with more than 25% missing. This code calculates the proportion missing from our variables and places them in rank order.

```
p_missing <- unlist(lapply(mimp_df, function(x) sum(is.na(x))))/nrow(mimp_df)
sort(p_missing[p_missing > 0], decreasing = TRUE)
```

	Blst_1	Blst_4	Blst_3	Blst_5	Blst_6	Belong_1	Belong_3
0.13043478	0.10144928	0.08695652	0.08695652	0.08695652	0.07246377	0.07246377	
Blst_2	Belong_2	cEval_1	cEval_19	iBIPOC_pr	cmBlack	cEval_13	
0.07246377	0.05797101	0.05797101	0.05797101	0.04347826	0.04347826	0.04347826	

Luckily, none of our variables have more than 25% missing. If we did have a variable with more than 25% missing, we would have to consider what to do about it.

Later we learn that we should eliminate case with greater than 50% missingness. Let's write code for that, now.

```
#Calculating number and proportion of item-level missingness
mimp_df$nmiss <- mimp_df %>%
  dplyr::select(iBIPOC_pr:format) %>% #the colon allows us to include all variables between
  is.na %>%
  rowSums

mimp_df <- mimp_df %>%
  dplyr::mutate(prop_miss = (nmiss/15)*100) #11 is the number of variables included in calculation

mimp_df <- dplyr::filter(mimp_df, prop_miss <= 50) #update df to have only those with at least
```

Once again, trim the df to include only the data to be included in the imputation

```
mimp_df <- select(mimp_df, ID, iBIPOC_pr, cmBlack, Belong_1:Belong_3, Blst_1:Blst_6,
  cEval_1, cEval_13, cEval_19, format)
```

### 4.5.3 The Multiple Imputation

Because multiple imputation is a *random* process, if we all want the same answers we need to set a *random seed*.

```
set.seed(210404) #you can pick any number you want, today I'm using today's timestamp
```

The program we will use is *mice*. *mice* assumes that each variable has a distribution and it imputes missing variables according to that distribution.

This means we need to correctly specify each variable's format/role. *mice* will automatically choose a distribution (think "format") for each variable; we can override this by changing the methods' characteristics.

The following code sets up the structure for the imputation. I'm not an expert at this – just following the Katitas example.

```
library(mice)
# runs the mice code with 0 iterations
imp <- mice(mimp_df, maxit = 0)
# Extract predictor Matrix and methods of imputation
predM = imp$predictorMatrix
meth = imp$method
```

Here we code what format/role each variable should be.

```

# These variables are left in the dataset, but setting them = 0 means
# they are not used as predictors. We want our ID to be retained in
# the df. There's nothing missing from it, and we don't want it used
# as a predictor, so it will just hang out.
predM[, c("ID")] = 0

# If you like, view the first few rows of the predictor matrix
# head(predM)

# We don't have any ordered categorical variables, but if we did we
# would follow this format poly <- c('Var1', 'Var2')

# We don't have any dichotomous variables, but if we did we would
# follow this format log <- c('Var3', 'Var4')

# Unordered categorical variables (nominal variables), but if we did
# we would follow this format
poly2 <- c("format")

# Turn their methods matrix into the specified imputation models
# Remove the hashtag if you have any of these variables meth[poly] =
# 'polr' meth[log] = 'logreg'
meth[poly2] = "polyreg"

meth

```

ID	iBIPOC_pr	cmBlack	Belong_1	Belong_2	Belong_3	Blst_1	Blst_2
" "	"pmm"	" "	"pmm"	" "	"pmm"	"pmm"	"pmm"
Blst_3	Blst_4	Blst_5	Blst_6	cEval_1	cEval_13	cEval_19	format
"pmm"	"pmm"	"pmm"	"pmm"	"pmm"	" "	"pmm"	"polyreg"

This list (meth) contains all our variables; “pmm” is the default and is the “predictive mean matching” process used. We see that format (an unordered categorical variable) is noted as “polyreg.” If we had used other categorical variables (ordered/poly, dichotomous/log), we would have seen those designations, instead. If there is “ ” underneath it means the data is complete.

Our variables of interest are now configured to be imputed with the imputation method we specified. Empty cells in the method matrix mean that those variables aren’t going to be imputed.

If a variable has no missing values, it is automatically set to be empty. We can also manually set variables to not be imputed with the *meth[variable]=“ ”* command.

The code below begins the imputation process. We are asking for 5 datasets. If you have many cases and many variables, this can take awhile. How many imputations? Recommendations have ranged as low as five to several hundred.

```

# With this command, we tell mice to impute the mimp_df data, create
# 5 datasets, use predM as the predictor matrix and don't print the

```

```
# imputation process. If you would like to see the process (or if
# the process is failing to execute) set print as TRUE; seeing where
# the execution halts can point to problematic variables (more notes
# at end of lecture)

imp2 <- mice(mimp_df, maxit = 5, predictorMatrix = predM, method = meth,
              print = FALSE)
```

We need to create a “long file” that stacks all the imputed data. Looking at the df in R Studio shows us that when imp = 0 (the pre-imputed data), there is still missingness. As we scroll through the remaining imputations, there are no NA cells.

```
# First, turn the datasets into long format This procedure is, best I
# can tell, unique to mice and wouldn't work for repeated measures
# designs
mimp_long <- mice::complete(imp2, action = "long", include = TRUE)
```

If we look at it, we can see 6 sets of data. If the *ID* variable is sorted we see that:

- .imp = 0 is the unimputed set; there are still missing values
- .imp = 1, 2, 3, or 5 has no missing values for the variables we included in the imputation

With the code below we can see the proportion of missingness for each variable (that has missing data), sorted from highest to lowest.

```
p_missing_mimp_long <- unlist(lapply(mimp_long, function(x) sum(is.na(x))))/nrow(mimp_long)
sort(p_missing_mimp_long[p_missing_mimp_long > 0], decreasing = TRUE) #check to see if this w
```

Blst_1	Blst_4	iBIPOC_pr	Blst_3	Blst_5	Blst_6
0.012820513	0.007692308	0.005128205	0.005128205	0.005128205	0.005128205
Belong_1	Belong_3	Blst_2	cEval_1	cEval_19	
0.002564103	0.002564103	0.002564103	0.002564103	0.002564103	

#### 4.5.4 Creating Scale Scores

Because our imputation was item-level, we need to score the variables with scales/subscales. As demonstrated more completely in the [Scoring chapter](#), this required reversing one item in the campus climate scale:

```
mimp_long <- mimp_long %>%
  mutate(rBlst_1 = 8 - Blst_1) #if you had multiple items, you could add a pipe (%>%) at th
```

Below is the scoring protocol we used in the AIA protocol for scoring. Although the protocol below functionally says, “Create a mean score if (65-80)% is non-missing, for the imputed version, it doesn’t harm anything to leave this because there is no missing data.

```
# Making the list of variables
Belonging_vars <- c("Belong_1", "Belong_2", "Belong_3")
ResponseBL_vars <- c("rBlst_1", "Blst_4", "Blst_6")
StigmaBL_vars <- c("Blst_2", "Blst_3", "Blst_5")
ClimateBL_vars <- c("rBlst_1", "Blst_4", "Blst_6", "Blst_2", "Blst_3",
"Blst_5")

# Creating the new variables
mimp_long$Belonging <- sjstats::mean_n(mimp_long[, Belonging_vars], 0.65)
mimp_long$ResponseBL <- sjstats::mean_n(mimp_long[, ResponseBL_vars], 0.8)
mimp_long$StigmaBL <- sjstats::mean_n(mimp_long[, StigmaBL_vars], 0.8)
mimp_long$ClimateBL <- sjstats::mean_n(mimp_long[, ClimateBL_vars], 0.8)
```

## 4.6 Multiple Regression with Multiply Imputed Data

For a refresher, here was the script when we used the AIA approach for managing missingness:

```
Climate_fit <- lm(ClimateBL ~ Belonging + cmBlack + iBIPOC_pr, data = item_scores_df)
summary(Climate_fit)
```

In order for the regression to use multiply imputed data, it must be a “mids” (multiply imputed data sets) type

```
# Convert to mids type - mice can work with this type
mimp_mids <- as.mids(mimp_long)
```

Here's what we do with imputed data:

```
fitimp <- with(mimp_mids, lm(ClimateBL ~ Belonging + cmBlack + iBIPOC_pr))
```

In this process, 5 individual, OLS, regressions are being conducted and the results being pooled into this single set.

```
# to get the 5, individual imputations
summary(fitimp)
```

term	estimate	std.error	statistic	p.value	nobs
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
1 (Intercept)	3.02	0.435	6.95	0.00000000283	65
2 Belonging	-0.0311	0.0897	-0.346	0.730	65
3 cmBlack	-0.0206	0.0165	-1.25	0.215	65
4 iBIPOC_pr	-0.663	0.339	-1.95	0.0552	65
5 (Intercept)	3.02	0.446	6.77	0.00000000578	65

6	Belonging	-0.0349	0.0907	-0.385	0.702	65
7	cmBlack	-0.0234	0.0166	-1.41	0.165	65
8	iBIPOC_pr	-0.470	0.329	-1.43	0.158	65
9	(Intercept)	3.01	0.450	6.70	0.00000000744	65
10	Belonging	-0.0349	0.0915	-0.381	0.704	65
11	cmBlack	-0.0222	0.0167	-1.33	0.187	65
12	iBIPOC_pr	-0.485	0.330	-1.47	0.147	65
13	(Intercept)	2.95	0.448	6.57	0.0000000127	65
14	Belonging	-0.0152	0.0920	-0.165	0.870	65
15	cmBlack	-0.0216	0.0168	-1.29	0.203	65
16	iBIPOC_pr	-0.558	0.343	-1.62	0.110	65
17	(Intercept)	3.00	0.452	6.64	0.00000000963	65
18	Belonging	-0.0311	0.0921	-0.337	0.737	65
19	cmBlack	-0.0214	0.0168	-1.28	0.207	65
20	iBIPOC_pr	-0.531	0.337	-1.57	0.121	65

```
pool(fitimp)
```

	Class:	mipo	m = 5	term	m	estimate	ubar	b	t	dfcom
1	(Intercept)	5	2.99980658	0.1990231323	0.000999315858	0.2002223113				61
2	Belonging	5	-0.02940746	0.0083160161	0.000067017541	0.0083964371				61
3	cmBlack	5	-0.02184241	0.0002777582	0.000001056566	0.0002790261				61
4	iBIPOC_pr	5	-0.54138195	0.1128248694	0.005817914953	0.1198063673				61
				df	riv	lambda	fmi			
1	58.70890	0.006025325	0.005989238	0.03820536						
2	58.44929	0.009670622	0.009577997	0.04181342						
3	58.80737	0.004564689	0.004543947	0.03675551						
4	53.13966	0.061879070	0.058273179	0.09182261						

```
summary(pool(fitimp))
```

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	2.99980658	0.44746208	6.7040465	58.70890	0.000000008735881
2	Belonging	-0.02940746	0.09163207	-0.3209298	58.44929	0.749408305666738
3	cmBlack	-0.02184241	0.01670407	-1.3076097	58.80737	0.196094825405891
4	iBIPOC_pr	-0.54138195	0.34613056	-1.5640975	53.13966	0.123730969370680

Results of a multiple regression predicting the respondents' perceptions of campus climate for Black students indicated that neither contributions of the respondents' personal belonging ( $B = -0.029, p = 0.749$ ), the proportion of BIPOC instructional staff ( $B = -0.541, p = 0.124$ ), nor proportion of Black classmates ( $B = -0.022, p = 0.196$ ) led to statistically significant changes in perceptions of campus climate for Black students. Results are presented in Table X.

## 4.7 Toward the APA Style Write-up

### 4.7.1 Method/Data Diagnostics

Data screening suggested that 107 individuals opened the survey link. Of those, 83 granted consent and proceeded into the survey items. A further inclusion criteria was that the course was taught in the U.S; 69 met this criteria.

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 67%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data. At this stage in the analysis, we allowed all cases with fewer than 50% missing to be included the multiple imputation [Katitas, 2019].

Regarding the distributional characteristics of the data, skew and kurtosis values of the variables fell below the values of 3 (skew) and 10 (kurtosis) that Kline suggests are concerning [2016b]. Results of the Shapiro-Wilk test of normality indicate that our variables assessing the proportion of classmates who are Black ( $W = 0.878, p < 0.001$ ) and the proportion of BIPOC instructional staff( $W = 0.787, p < 0.001$ ) are statistically significantly different than a normal distribution. The scales assessing the respondent's belonging (0.973,  $p = 0.165$ ) and the respondent's perception of campus climate for Black students ( $W = 0.951, p = 0.016$ ) did not differ differently from a normal distribution.

We evaluated multivariate normality with the Mahalanobis distance test. Specifically, we used the *psych::outlier()* function and included all continuous variables in the calculation. Our visual inspection of the Q-Q plot suggested that the plotted line strayed from the straight line as the quantiles increased. Additionally, we appended the Mahalanobis distance scores as a variable to the data. Analyzing this variable, we found that 1 case exceed three standard deviations beyond the median. Given that the Mahalanobis distance values increased in a consistent manner (i.e., no extreme “jumps”) we retained all cases.

We managed missing data with multiple imputation [Enders, 2017, Katitas, 2019]. We imputed five sets of data with the R package, *mice* (v. 3.13) – a program that utilizes conditional multiple imputation. The imputation included the item-level variables that comprised our scales, the variables that represented proportion of BIPOC instructional staff and proportion of Black classmates, as well as four auxiliary variables (three variables from the course evaluation and the format [in-person/blended/virtual] of the class).

### 4.7.2 Results

Results of a multiple regression predicting the respondents' perceptions of campus climate for Black students indicated that neither contributions of the respondents' personal belonging ( $B = -0.029, p = 0.749$ ), *the proportion of BIPOC instructional staff* ( $B = -0.541, p = 0.124$ ), nor proportion of Black classmates ( $B = -0.022, p = 0.196$ ) led to statistically significant changes in perceptions of campus climate for Black students. Results are presented in Table X.

#### Some notes about this write-up

- I went ahead and used the data diagnostics that we did in the AIA method. It feels to me like these should be calculated with the multiply imputed data (i.e., 5 sets, with pooled estimates and standard errors), but I do not see that modeled – anywhere in R.
- Note the similarities with the AIA write-up.

## 4.8 Multiple imputation considerations

- Character vectors (i.e., values that are represented with words) can be problematic. If they are causing trouble, consider
  - recode into factors,
  - keep it in the df, but exclude it from the imputation protocol,
  - our “format” variable was an ordered factor (i.e., each term was associated with a value), so I think that helped us avoid problems
- Variables with really high (like 50% or more) proportions of missingness should be excluded.
- Variables that are highly correlated or redundant (even if inverse) will halt the execution. If you set `print=TRUE` you will see where the algorithm is having difficulty because it will halt at that variable.
- Variables with non-missing values can be problematic. If they are problematic, just exclude them from the process. \*Width (columns/variables) versus length (rows/cases). You must have more rows/cases than columns/variables. It is difficult to say how many. If this is a problem:
  - Consider scoring scales first with AIA, then impute with whole scales.
  - Divide the df in halves or thirds, impute separately, then join with the ID numbers.
  - There should be auxiliary variables in each. \*Item-level imputation is its “whole big thing” with multiple, complex considerations. There are tremendous resources
  - Enders [BLIMP](#) app is free and works with R
  - Little’s [2002] article
- How many imputations? Controversial and has changed over the years.
  - Practical concern: the more you request, the longer it will take in R, this demo was 5
  - For a number of years there was a push for 20, but I’ve also seen recommendations for 100s.

- Check examples of imputed studies in your disciplinary specialty/journals.
- There are lots of discussions and debates about
  - allowing for fractional/decimal responses (a 3.5 on 1 to 4 scaling; or a 0.75 on a dichotomous variable such as male/female)
  - out-of-bounds estimates (what if you get a 7 on 1 to 4 scaling?)

## 4.9 Practice Problems

The three problems described below are designed to be continuations from the previous chapters. You will likely encounter challenges that were not covered in this chapter. Search for and try out solutions, knowing that there are multiple paths through the analysis. In addition to the scrubbing, scoring, and data diagnostic skills learned in the prior lessons, the overall notion of the suggestions for practice are to (a) multiply impute a minimum of 5 sets of data, (b) repeat the regression (attempted in the Data Dx chapter), (c) create APA style write-ups of the multiple imputation method and regression results, and (d) explain it to someone.

### 4.9.1 Problem #1: Reworking the Chapter Problem

If you chose this option in the prior chapters, you imported the data from Qualtrics, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, properly formatted the variables, interpreted item-level missingness, scored the scales/subscales, interpreted scale-level missingness, and wrote up the results. Please continue with the remaining tasks.

### 4.9.2 Problem #2: Use the *Rate-a-Recent-Course* Survey, Choosing Different Variables

If you chose this option in the prior chapter, you chose a minimum of three variables from the *Rate-a-Recent-Course* survey to include in a simple statistical model. You imported the data from Qualtrics, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, properly formatted the variables, interpreted item-level missingness, scored the scales/subscales, interpreted scale-level missingness, and wrote up the results. Please continue with the remaining tasks.

### 4.9.3 Problem #3: Other data

If you chose this option in the prior chapter, you used raw data that was available to you. You imported it into R, applied inclusion/exclusion criteria, renamed variables, downsized the df to the variables of interest, properly formatted the variables, interpreted item-level missingness, scored the scales/subscales, interpreted scale-level missingness, and wrote up the results. Please continue with the remaining tasks.

#### 4.9.4 Grading Rubric

Assignment Component	Points Possible	Points Earned
1. Specify a research model with three predictor variables (continuously or categorically scaled) and one dependent (continuously scaled) variable.	5	_____
2. Import the raw data	5	_____
3. Apply inclusionary/exclusionary criteria	5	_____
4. Format any variables that shouldn't be imputed in their raw form	5	_____
5. Multiply impute a minimum of 5 sets of data	5	_____
6. Run a regression (for multiply imputed data) with at least three variables	5	_____
7. APA style write-up of the multiple imputation section of data diagnostics	5	_____
8. APA style write-up regression results	5	_____
9. Explanation to grader	5	_____
<b>Totals</b>	<b>45</b>	_____

## 4.10 Homeworked Example

Screencast Link

For more information about the data used in this homeworked example, please refer to the description and codebook located at the end of the [introductory lesson](#) in [ReCentering Psych Stats](#). An .rds file which holds the data is located in the [Worked Examples](#) folder at the GitHub site the hosts the OER. The file name is *ReC.rds*.

Although the lessons focused on preparing data for analyses were presented in smaller sections, this homeworked example combines the suggestions for practice from the [Scrubbing](#), [Scoring](#), [Data Dx](#) because they are also used when missing data is managed with multiple imputation. My hope is that is cumulative presentation is a closer approximation of what researchers need for their research projects.

These lessons were created to prepare a set of data to analyze a specific research model. Consequently, the model should be known and described at the beginning.

### 4.10.1 Scrubbing

#### Specify a research model

A further assignment requirement was that the model should include three predictor variables (continuously or categorically scaled) and one dependent (continuously scaled) variable.

As in the homeworked example for the Data Dx lesson, I am hypothesizing that socially responsive pedagogy (my dependent variable) will increase as a function of:

- the transition from SPSS (0) to R(1),
- the transition from a pre-centered (0) to re-centered (1) curriculum, and
- higher evaluations of traditional pedagogy

Because this data is nested within the person (i.e., students can contribute up to three course evaluations over the ANOVA, multivariate, and psychometrics courses) proper analysis would require a statistic (e.g., multilevel modeling) that would address the dependency in the data. Therefore, I will include only those students who are taking the multivariate modeling class.

While it is possible to conduct multiple imputation at the scale level, we will do so at the item-level (i.e., before we compute the scale scores).

*If you wanted to use this example and dataset as a basis for a homework assignment, you could create a different subset of data. I worked the example for students taking the multivariate modeling class. You could choose ANOVA or psychometrics. You could also choose a different combinations of variables.*

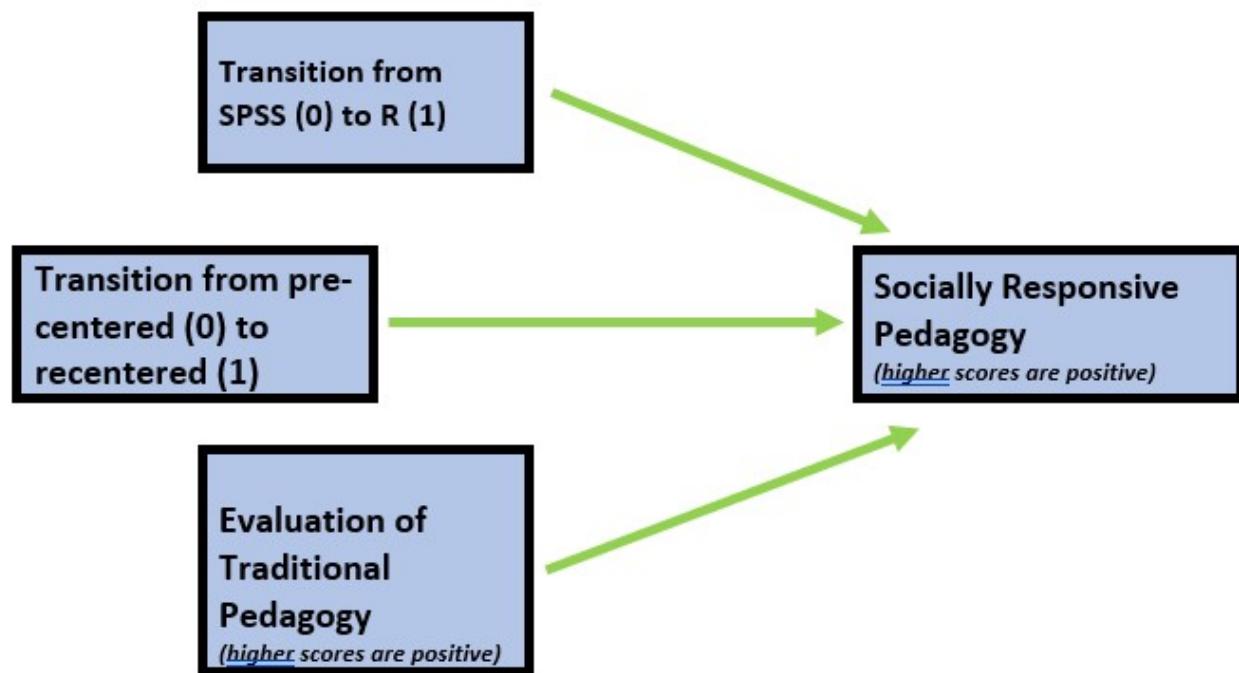


Figure 4.5: An image of our the prediction model for the homeworked example.

### Import data

```
raw <- readRDS("ReC.rds")
nrow(raw)
```

[1] 310

## Apply inclusionary/exclusionary criteria

Because this data is publicly posted on the Open Science Framework, it was necessary for me to already exclude those individuals. This data was unique in that students could freely write some version of “Opt out.” My original code included a handful of versions, but here was the basic form:

```
# testing to see if my code worked raw <- dplyr::filter (raw,
# SPFC.Decolonize.Opt.Out != 'Okay')
raw <- dplyr::filter(raw, SPFC.Decolonize.Opt.Out != "Opt Out")
```

I want to exclude students’ responses for the ANOVA and psychometrics courses.

```
raw <- dplyr::filter(raw, Course == "Multivariate")
```

At this point, these my only inclusion/exclusion criteria. I can determine how many students (who consented) completed any portion of the survey.

```
nrow(raw)
```

```
[1] 84
```

### 4.10.1.1 Format any variables that shouldn’t be imputed in their raw form

Let’s first create a df with the item-level variables that will fuel our model.

In addition to the variables in our model, we will include four auxiliary variables. These include Dept (Department: Clinical or Industrial-Organizational) and four additional course evaluation items: OvInstructor, MyContribution, IncrInterest, IncrUnderstanding.

Let’s check the structure to be certain that *StatsPkg* (SPSS, R) and *Centered* (Pre, Re) are ordered factors. We also want the course evaluation items to be integer (or numerical).

```
mimp_df <- dplyr::select(raw, deID, StatsPkg, Centering, ClearResponsibilities,
  EffectiveAnswers, Feedback, ClearOrganization, ClearPresentation, InclusvClassrm,
  EquitableEval, MultPerspectives, DEIIntegration, Dept, OvInstructor,
  MyContribution, IncrInterest, IncrUnderstanding)
str(mimp_df)
```

```
Classes 'data.table' and 'data.frame': 84 obs. of 17 variables:
 $ deID           : int 11 12 13 14 15 16 17 18 35 19 ...
 $ StatsPkg       : Factor w/ 2 levels "SPSS","R": 2 2 2 2 2 2 2 2 2 2 ...
 $ Centering      : Factor w/ 2 levels "Pre","Re": 2 2 2 2 2 2 2 2 2 2 ...
 $ ClearResponsibilities: int 4 5 5 5 4 3 5 5 3 5 ...
 $ EffectiveAnswers : int 4 5 5 4 4 3 5 5 4 4 ...
 $ Feedback        : int 4 5 4 4 5 4 5 4 4 5 ...
 $ ClearOrganization: int 3 5 5 4 4 3 5 5 4 5 ...
```

```
$ ClearPresentation      : int  4 5 5 3 4 2 5 4 5 5 ...
$ InclusvClassrm       : int  5 5 5 5 4 5 5 5 5 ...
$ EquitableEval         : int  4 5 5 5 4 4 5 4 5 5 ...
$ MultPerspectives      : int  4 5 5 5 5 5 5 4 5 5 ...
$ DEIintegration        : int  5 5 5 5 5 5 5 5 5 5 ...
$ Dept                  : chr  "CPY" "CPY" "CPY" "CPY" ...
$ OvInstructor          : int  3 5 5 3 5 2 5 4 5 5 ...
$ MyContribution        : int  4 5 4 3 4 3 5 4 4 5 ...
$ IncrInterest          : int  4 5 4 3 4 3 5 4 5 4 ...
$ IncrUnderstanding     : int  4 5 5 3 4 3 5 4 5 5 ...
- attr(*, ".internal.selfref")=<externalptr>
```

```
mimp_df$Dept <- factor(mimp_df$Dept, levels = c("CPY", "ORG"))
str(mimp_df$Dept)
```

Factor w/ 2 levels "CPY","ORG": 1 1 1 1 1 1 1 1 1 1 ...

We should eliminate case with greater than 50% missingness.

```
library(tidyverse)
#Calculating number and proportion of item-level missingness
mimp_df$nmiss <- mimp_df%>%
  dplyr::select(StatsPkg:IncrUnderstanding) %>% #the colon allows us to include all variables
  is.na %>%
  rowSums

mimp_df<- mimp_df%>%
  dplyr::mutate(prop_miss = (nmiss/13)*100) #11 is the number of variables included in calculation

mimp_df <- filter(mimp_df, prop_miss <= 50) #update df to have only those with at least 50% o
```

Once again, trim the df to include only the data to be included in the imputation

```
mimp_df <- dplyr::select(mimp_df, deID, StatsPkg, Centering,ClearResponsibilities, EffectiveA
```

#### 4.10.1.2 Multiply impute a minimum of 5 sets of data

Because multiple imputation is a *random* process, if we all want the same answers we need to set a *random seed*.

```
set.seed(2309034) #you can pick any number you want, today I'm using today's timestamp
```

The program we will use is *mice*. *mice* assumes that each variable has a distribution and it imputes missing variables according to that distribution.

This means we need to correctly specify each variable's format/role. *mice* will automatically choose a distribution (think “format”) for each variable; we can override this by changing the methods' characteristics.

The following code sets up the structure for the imputation. This follows the Katitas example.

```
library(mice)
# runs the mice code with 0 iterations
imp <- mice(mimp_df, maxit = 0)
# Extract predictor Matrix and methods of imputation
predM = imp$predictorMatrix
meth = imp$method
log = imp$log
```

Here we code what format/role each variable should be.

```
# These variables are left in the dataset, but setting them = 0 means
# they are not used as predictors. We want our ID to be retained in
# the df. There's nothing missing from it, and we don't want it used
# as a predictor, so it will just hang out.
predM[, c("deID")] = 0

# If you like, view the first few rows of the predictor matrix
# head(predM)

# We don't have any ordered categorical variables, but if we did we
# would follow this format poly <- c('Var1', 'Var2')

# We have three dichotomous variables
log <- c("StatsPkg", "Centering", "Dept")

# Unordered categorical variables (nominal variables), but if we did
# we would follow this format poly2 <- c('format')

# Turn their methods matrix into the specified imputation models
# Remove the hashtag if you have any of these variables meth[poly] =
# 'polr'
meth[log] = "logreg"
# meth[poly2] = 'polyreg'

meth
```

	StatsPkg	Centering
deID	"logreg"	"logreg"
ClearResponsibilities	EffectiveAnswers	Feedback
"pmm"	"	"pmm"
ClearOrganization	ClearPresentation	InclusvClassrm
""	""	"pmm"

EquitableEval	MultPerspectives	DEIintegration
" "	"pmm"	"pmm"
Dept	OvInstructor	MyContribution
"logreg"	" "	" "
IncrInterest	IncrUnderstanding	" "
"pmm"		

This list (*meth*) contains all our variables; “pmm” is the default and is the “predictive mean matching” process used. We see that *StatsPkg* and *Centering* are noted as “logreg.” This is because they are dichotomous variables. If there is “ ” underneath it means the data is complete. The data will be used in imputing other data, but none of that data will be imputed.

Our variables of interest are now configured to be imputed with the imputation method we specified. Empty cells in the method matrix mean that those variables aren’t going to be imputed.

If a variable has no missing values, it is automatically set to be empty. We can also manually set variables to not be imputed with the *meth[variable] = “ ”* command.

The code below begins the imputation process. We are asking for 5 datasets. If you have many cases and many variables, this can take awhile. How many imputations? Recommendations have ranged as low as five to several hundred.

```
# With this command, we tell mice to impute the anesimpor2 data,
# create 5vudatasets, use predM as the predictor matrix and don't
# print the imputation process. If you would like to see the process
# (or if the process is failing to execute) set print as TRUE; seeing
# where the execution halts can point to problematic variables (more
# notes at end of lecture)

imp2 <- mice(mimp_df, maxit = 5, predictorMatrix = predM, method = meth,
  log = log, print = FALSE)
```

We need to create a “long file” that stacks all the imputed data. Looking at the df in R Studio shows us that when *imp* = 0 (the pre-imputed data), there is still missingness. As we scroll through the remaining imputations, there are no NA cells.

```
# First, turn the datasets into long format This procedure is, best I
# can tell, unique to mice and wouldn't work for repeated measures
# designs
mimp_long <- mice::complete(imp2, action = "long", include = TRUE)
```

If we look at it, we can see 6 sets of data. If the *deID* variable is sorted we see that:

- *.imp* = 0 is the unimputed set; there are still missing values
- *.imp* = 1, 2, 3, or 5 has no missing values for the variables we included in the imputation

With the code below we can see the proportion of missingness for each variable (that has missing data), sorted from highest to lowest.

```
p_missing_mimp_long <- unlist(lapply(mimp_long, function(x) sum(is.na(x))))/nrow(mimp_long)
sort(p_missing_mimp_long[p_missing_mimp_long > 0], decreasing = TRUE) #check to see if this works
```

DEIIntegration	InclusvClassrm	Feedback
0.027777778	0.007936508	0.003968254

ClearResponsibilities	MultPerspectives	IncrInterest
0.001984127	0.001984127	0.001984127

Because our imputation was item-level, we need to score the variables with scales/subscales.

Traditional pedagogy is a predictor variable that needs to be created by calculating the mean if at least 75% of the items are non-missing. None of the items need to be reverse-scored. I will return to working with the *scrub\_df* data.

```
# this seems to work when I build the book, but not in 'working the
# problem'
TradPed_vars <- c("ClearResponsibilities", "EffectiveAnswers", "Feedback",
  "ClearOrganization", "ClearPresentation")
# mimp_long$TradPed <- sjstats::mean_n(mimp_long[, TradPed_vars],
# .75)

# this seems to work when I 'work the problem' (but not when I build
# the book) the difference is the two dots before the last SRPed_vars
mimp_long$TradPed <- sjstats::mean_n(mimp_long[, TradPed_vars], 0.75)
```

The dependent variable is socially responsive pedagogy. It needs to be created by calculating the mean if at least 75% of the items are non-missing. None of the items need to be reverse-scored.

```
# this seems to work when I build the book, but not in 'working the
# problem' SRPed_vars <- c('InclusvClassrm', 'EquitableEval',
# 'MultPerspectives', 'DEIIntegration') mimp_long$SRPed <-
# sjstats::mean_n(mimp_long[, SRPed_vars], .75)

# this seems to work when I 'work the problem' (but not when I build
# the book) the difference is the two dots before the last SRPed_vars
SRPed_vars <- c("InclusvClassrm", "EquitableEval", "MultPerspectives",
  "DEIIntegration")
mimp_long$SRPed <- sjstats::mean_n(mimp_long[, SRPed_vars], 0.75)
```

#### 4.10.1.3 Run a regression (for multiply imputed data) with at least three variables

For comparison, here was the script when we used the AIA approach for managing missingness:

```
SRPed_fit <- lm(SRPed ~ StatsPkg + Centering + TradPed, data = scored)
```

In order for the regression to use multiply imputed data, it must be a “mids” (multiply imputed data sets) type

```
# Convert to mids type - mice can work with this type
mimp_mids <- as.mids(mimp_long)
```

Here's what we do with imputed data:

```
fitimp <- with(mimp_mids, lm(SRPed ~ StatsPkg + Centering + TradPed))
```

In this process, 5 individual, OLS, regressions are being conducted and the results being pooled into this single set.

```
# to get the 5, individual imputations
summary(fitimp)
```

	term	estimate	std.error	statistic	p.value	nobs
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
1	(Intercept)	1.90	0.310	6.13	3.13e- 8	84
2	StatsPkgR	0.187	0.118	1.59	1.16e- 1	84
3	CenteringRe	0.117	0.108	1.09	2.79e- 1	84
4	TradPed	0.565	0.0659	8.56	6.30e-13	84
5	(Intercept)	1.94	0.314	6.17	2.62e- 8	84
6	StatsPkgR	0.191	0.119	1.60	1.13e- 1	84
7	CenteringRe	0.110	0.109	1.01	3.17e- 1	84
8	TradPed	0.557	0.0667	8.36	1.63e-12	84
9	(Intercept)	1.96	0.313	6.26	1.80e- 8	84
10	StatsPkgR	0.178	0.119	1.50	1.38e- 1	84
11	CenteringRe	0.111	0.109	1.02	3.10e- 1	84
12	TradPed	0.555	0.0665	8.35	1.69e-12	84
13	(Intercept)	2.03	0.325	6.24	1.98e- 8	84
14	StatsPkgR	0.185	0.123	1.50	1.38e- 1	84
15	CenteringRe	0.104	0.113	0.918	3.62e- 1	84
16	TradPed	0.539	0.0691	7.80	1.95e-11	84
17	(Intercept)	1.91	0.306	6.26	1.77e- 8	84
18	StatsPkgR	0.158	0.116	1.36	1.78e- 1	84
19	CenteringRe	0.117	0.107	1.10	2.76e- 1	84
20	TradPed	0.567	0.0649	8.73	2.93e-13	84

```
summary(pool(fitimp))
```

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	1.9480744	0.31833535	6.119567	74.55114	0.000000040039269753
2	StatsPkgR	0.1798400	0.11996611	1.499090	76.64577	0.137957984459613908
3	CenteringRe	0.1117906	0.10918108	1.023901	77.81162	0.309054914517060075
4	TradPed	0.5564494	0.06768356	8.221338	74.26455	0.000000000004825124

Results of a multiple regression predicting the socially responsive course evaluation ratings indicated that neither the transition from SPSS to R ( $B = 0.178, p = 0.135$ ) nor the transition to an explicitly recentered curriculum ( $B = 0.116, p = 0.285$ ) led to statistically significant differences. In contrast, traditional pedagogy had a  $B = 0.571, p < 0.001$ ). Results of the regression model are presented in Table 2.

#### 4.10.1.4 APA style write-up of the multiple imputation section of data diagnostics

My write-up draws from some of the results we obtained in the homeworked example at the end of the **Data Dx** chapter.

This is a secondary analysis of data involved in a more comprehensive dataset that included students taking multiple statistics courses ( $N = 310$ ). Having retrieved this data from a repository in the Open Science Framework, only those who consented to participation in the study were included. Data used in these analyses were 84 students who completed the multivariate clas.

Across cases that were deemed eligible on the basis of the inclusion/exclusion criteria, missingness ranged from 0 to 100%. Across the dataset, 3.86% of cells had missing data and 87.88% of cases had nonmissing data. At this stage in the analysis, missingness for all cases did not exceed 50% [Katitas, 2019] and they were all included in the multiple imputation .

Regarding the distributional characteristics of the data, skew and kurtosis values of the variables fell below the values of 3 (skew) and 10 (kurtosis) that Kline suggests are concerning [2016b]. Results of the Shapiro-Wilk test of normality indicate that our variables assessing the traditional pedagogy ( $W = 0.830, p < 0.001$ ) and socially responsive pedagogy (0.818,  $p < 0.001$ ) are statistically significantly different than a normal distribution. Inspection of distributions of the variables indicated that both course evaluation variables were negatively skewed, with a large proportion of high scores.

We evaluated multivariate normality with the Mahalanobis distance test. Specifically, we used the *psych::outlier()* function and included both continuous variables in the calculation. Our visual inspection of the Q-Q plot suggested that the plotted line strayed from the straight line as the quantiles increased. Additionally, we appended the Mahalanobis distance scores as a variable to the data. Analyzing this variable, we found that 2 cases exceed three standard deviations beyond the median.

We managed missing data with multiple imputation [Enders, 2017, Katitas, 2019]. We imputed five sets of data with the R package, *mice* (v. 3.13) – a program that utilizes conditional multiple imputation. The imputation included the 9 item-level variables that comprised our scales and the dichotomous variable representing traditional pedagogy and socially responsive pedagogy. We also included five auxiliary variables (four variables from the course

evaluation and the whether the student was from the Clinical or Industrial-Organizational Psychology program).

#### 4.10.1.5 APA style write-up regression results

Results of a multiple regression predicting the socially responsive course evaluation ratings indicated that neither the transition from SPSS to R ( $B = 0.178, p = 0.135$ ) nor the transition to an explicitly recentered curriculum ( $B = 0.116, p = 0.285$ ) led to statistically significant differences. In contrast, traditional pedagogy had a  $B = 0.571, p < 0.001$ ). Results of the regression model are presented in Table 2.

*As in the lesson itself, I used the data diagnostics that we did in the AIA method. It feels to me like these should be calculated with the multiply imputed data (i.e., 5 sets, with pooled estimates and standard errors), but I do not see that modeled – anywhere in tutorials I consulted.*



# MEDIATION



# Chapter 5

## Simple Mediation

### [Screencasted Lecture Link](#)

The focus of this lecture is to estimate indirect effects (aka “mediation”). We examine the logic/design required to support the argument that *mediation* is the *mechanism* that explains the  $X \rightarrow Y$  relationship. We also work three examples (one with covariates).

At the outset, please note that although I rely heavily on Hayes [2018] text and materials, I am using the R package *lavaan* in these chapters. In recent years, Hayes has introduced a [PROCESS macro for R](#). Because I am not yet up-to-speed on using this macro (it is not a typical R package) and because ReCentering Psych Stats uses *lavaan* for confirmatory factor analysis and structural equation modeling, I have chosen to utilize the *lavaan* package. A substantial difference is that the PROCESS macros use ordinary least squares and *lavaan* uses maximum likelihood estimators.

### 5.1 Navigating this Lesson

There is about 1 hour and 10 minutes of lecture. If you work through the materials with me it would be plan for an additional 1.5 hours.

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER’s [introduction](#)

#### 5.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Define mediation and indirect effect.
- Distinguish the role of a mediating variable from independent variables, covariates, and moderators.
- Identify the conditions upon which there can be justification to support the presence of a mediated effect.

- Articulate the arguments for and against using the term, “mediation.”
- Using the R package *lavaan*,
  - specify a model with indirect effects,
  - identify and interpret B weights, *p* values, and *CIs* for total, direct, and indirect effects,
  - calculate the total effects of X and M on Y,
  - identify the proportion of variance accounted for in predicting M and Y.
- Hand calculate the values of an indirect, direct, and total effects from statistical output or a figure (just the *B* or  $\beta$ , not the significance level)

### 5.1.2 Planning for Practice

The following suggestions for practice will involve specifying, testing, and interpreting a model with a single indirect effect (mediator).

- Rework the problem in the chapter by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- There are a number of variables in the dataset and there were a handful of simple mediations conducted in the journal article that sources the research vignette. Swap out one or more variables in the model of simple mediation and compare your solution to the one in the chapter and/or the research article.
- Conduct a simple mediation with data to which you have access. This could include data you simulate on your own or from a published article.

### 5.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press. Available as an ebook from the SPU library: <https://ebookcentral-proquest-com.ezproxy.spu.edu/lib/spu/detail.action?docID=5109647>
  - **Chapter 3, Simple mediation:** Hayes’ text is another great example of a teaching tool that is accessible at both procedural and conceptual levels. I especially appreciate his attention to the controversies (even those directed toward his work). We deviate from his text in that we are not using the PROCESS macro...and I’ll address those concerns in the lecture.
  - **Chapter 4, Causality and confounds:** A great chapter that addresses “What happened to Baron & Kenny”; partial v complete mediation; and conditions required for claims of causality. Procedurally, our focus in this chapter is on the role of covariates.
  - **Appendix A: Using Process:** An essential tool for PROCESS users because, even when we are in the R environment, this is the “idea book.” That is, the place where all the path models are presented in figures.

- Kim, P. Y., Kendall, D. L., & Cheon, H.-S. (2017). Racial microaggressions, cultural mistrust, and mental health outcomes among Asian American college students. *American Journal of Orthopsychiatry*, 87(6), 663–670. <https://doi-org.ezproxy.spu.edu/10.1037/ort0000203>

#### 5.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
if (!require(lavaan)) {
  install.packages("lavaan")
}
if (!require(semPlot)) {
  install.packages("semPlot")
}
if (!require(tidyverse)) {
  install.packages("tidyverse")
}
if (!require(psych)) {
  install.packages("psych")
}
if (!require(formattable)) {
  install.packages("formattable")
}
if (!require(semTable)) {
  install.packages("semTable")
}
```

## 5.2 Estimating Indirect Effects (the analytic approach often termed *mediation*)

### 5.2.1 The definitional and conceptual

As in Hayes text [2018], we will differentiate between *moderation* and *mediation*. *Conditional process analysis* involves both! With each of these, we are seeking to understand the *mechanism* at work that leads to the relationship (be it correlational, predictive, or causal)

Even though this process has sometimes been termed *causal modeling*, Hayes argues that his *statistical approach* is not claiming to determine *cause*; that is really left to the argument of the research design.

**Moderation** (a review):

- Answers questions of *when* or *for whom* and is often the source of the answer, *it depends*.
- Think of our *interaction* effects in ANOVA and regression

- The effect of X on some variable Y is moderated by W if its size, sign, or strength depends on, or can be predicted, by W. Then we can say, “W is a *moderator* of X’s effect on Y” or “W and X *interact* in their influence on Y.”
- The image below illustrates moderation with *conceptual* and *statistical* diagrams. Note that three predictors (IV, DV, their interaction) point to the DV.

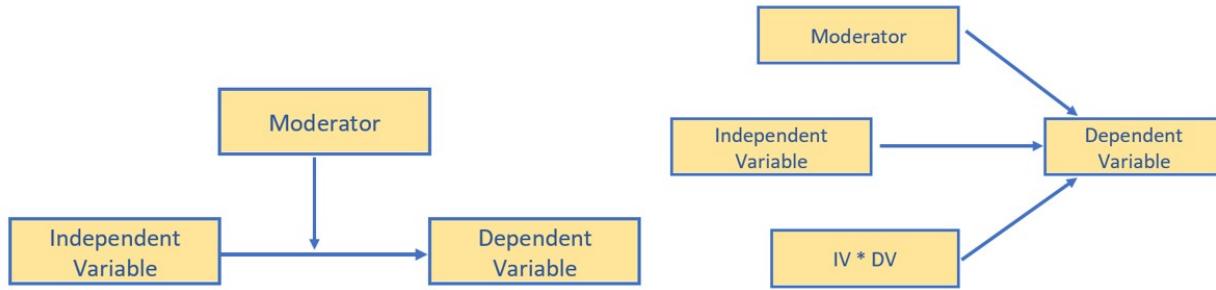


Figure 5.1: Image of Hayes's style conceptual and statistical diagrams of a simple moderation

The classic plot of moderation results is often the best way to detect that an interaction was included in the analysis and helps understand the *conditional* (e.g., for whom, under what conditions) nature of the analysis.

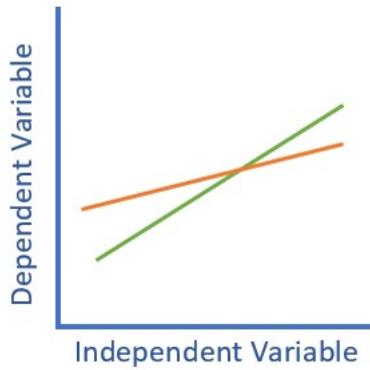


Figure 5.2: Image of classic interaction graph that illustrates a moderated effect. The IV is on the X axis, DV on the Y axis, and two intersecting lines represent the differential/moderated effect of the IV on the DV by the moderator

### Mediation:

- Answers questions of *how* (I also think *through* and *via* to describe the proposed mediating mechanism)
- Paths in a mediation model are *direct* (X does not pass through M on its way to Y) and *indirect* (X passes through M on its way to Y). Once we get into the statistics, we will also be focused on *total* effects.

- Hayes thinks in terms of *antecedent* and *consequent* variables. In a 3-variable, simple mediation, X and M are the antecedent variables; X and M are the consequent variables.
- There is substantial debate and controversy about whether we can say “the effect of X on Y is *mediated* through M” or whether we should say, “There is a statistically significant indirect effect of X on Y thru M.” Hayes comes down on the “use mediation language” side of the debate.
- In sum, a simple mediation model is any causal system in which at least one causal antecedent X variable is proposed as influencing an outcome Y through a single intervening variable, M. In such a model there are two pathways by which X can influence Y.
- The figure below doubles as both the conceptual and statistical diagram of evaluating a simple mediation – a simple indirect effect.

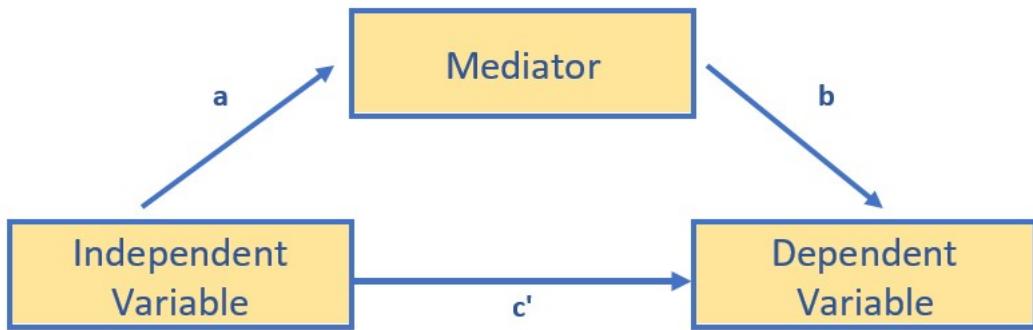


Figure 5.3: Image of Hayes's style conceptual diagram of a simple moderation

#### Conditional process analysis:

- Used when the research goal is to understand the boundary conditions of the mechanism(s) by which a variable transmits its effect on another.
- Typically, simultaneously, assesses the influence of mediating (indirect effects) and moderating (interactional effects) in a model-building fashion.
- In a conditional process model, the moderator(s) may be hypothesized to influence one or more of the paths.

We will work toward building a conditional process model, a moderated mediation, over the next several chapters.

### 5.3 Workflow for Simple Mediation

The following is a proposed workflow for conducting a simple mediation.

Conducting a simple mediation involves the following steps:

1. Conducting an a priori power analysis to determine the appropriate sample size.

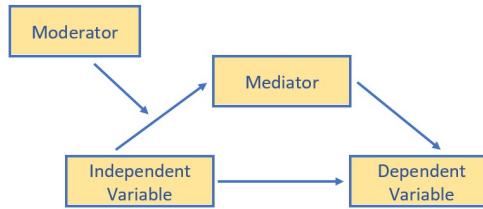


Figure 5.4: Image of conditional process analysis model where the moderator is hypothesized to change the a path; the path between the IV and mediator

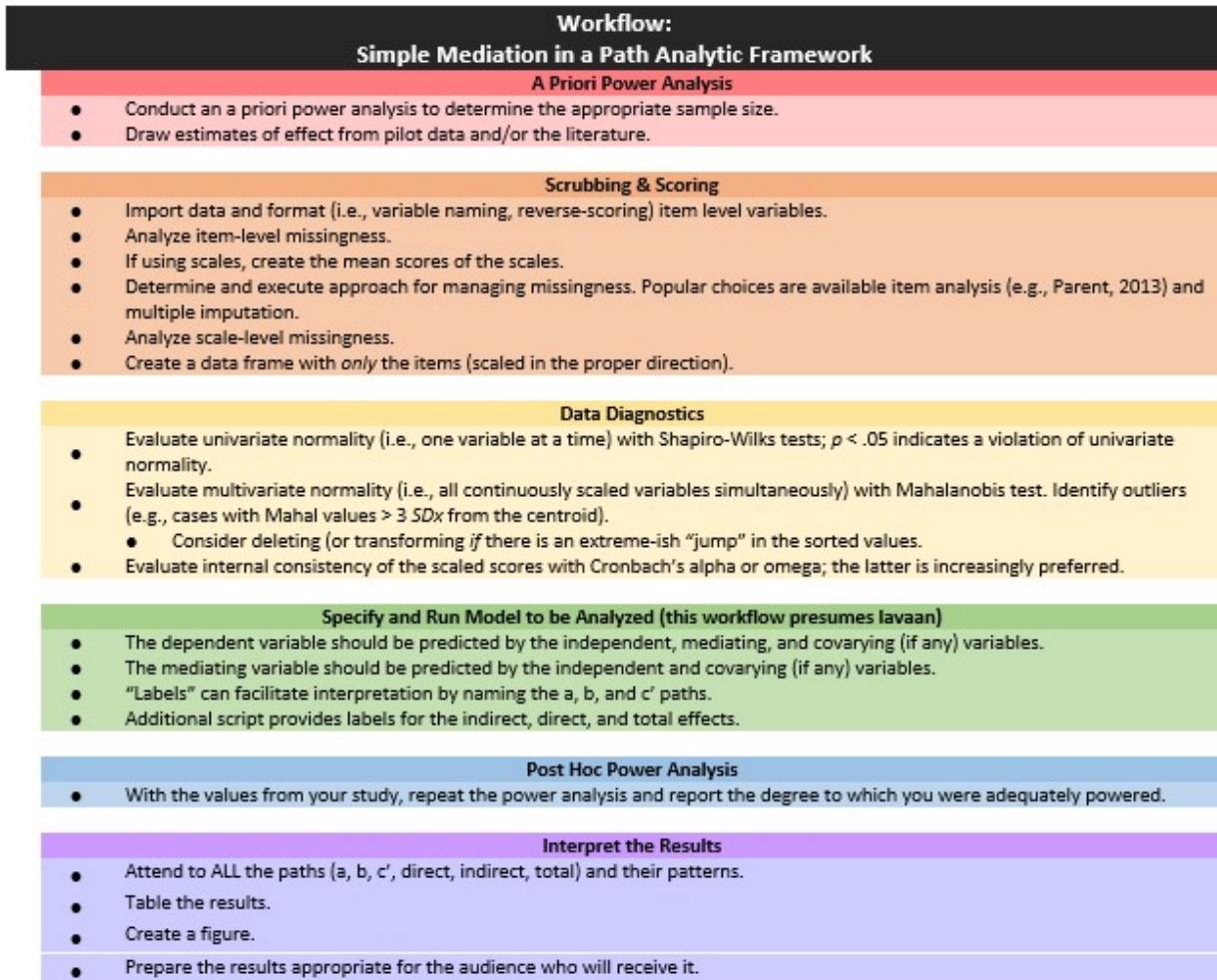


Figure 5.5: A colorful image of a workflow for the simple mediation

- This will require estimates of effect that are drawn from pilot data, the literature, or both.
2. Scrubbing and scoring the data.
    - Guidelines for such are presented in the respective lessons.
  3. Conducting data diagnostics, this includes:
    - item and scale level missingness,
    - internal consistency coefficients (e.g., alphas or omegas) for scale scores,
    - univariate and multivariate normality
  4. Specifying and running the model (this lesson presumes it will with the R package, *lavaan*).
    - The dependent variable should be predicted by the independent, mediating, and covarying (if any) variables.
    - “Labels” can facilitate interpretation by naming the  $a$ ,  $b$ , and  $c'$  paths. +Additional script provides labels for the indirect, direct, and total effects.
  5. Conducting a post hoc power analysis.
    - Informed by your own results, you can see if you were adequately powered to detect a statistically significant effect, if, in fact, one exists.
  6. Interpret and report the results.
    - Interpret ALL the paths and their patterns.
    - Create a table and figure.
    - Prepare the results in a manner that is useful to your audience.

In addition to the workflow through the statistical problem, the very traditional and classic figure below is useful in understanding the logic beneath mediation as the explanatory mechanism.

The top figure represents the bivariate relationship between the independent and dependent variable. The result of a simple linear regression (one predictor) represent the *total* effect of the IV on the DV. We can calculate this by simply regressing the DV onto the IV. The resulting  $B$  weight is known as the  $c$  path. A bivariate correlation coefficient results in the same value – only it is standardized (so would be the same as the  $\beta$  weight).

The lower figure represents that the relationship between the IV and DV is *mediated* by a third variable. We assign three labels to the paths:  $a$ , between the IV and mediator;  $b$ , between the mediator and DV; and  $c'$  ( $c$  prime) between the IV and DV.

Although Hayes makes a compelling case that we can claim “mediation” when there is a statistically significant indirect effect [2018], traditionally, a mediated relationship is supported when the value of  $c'$  is statistically significantly lower than  $c$ . When this occurs, then know that the mediator is sharing some of the variance (and therefore acting as a *conduit*) in the prediction of the DV.

You might already be imagining potential challenges to this model. For example, which variable should be the IV and which one should be the mediator? Can we switch them? You can – and you will likely have very similar (if not identical) results. Good research design is what provides support for suggesting that mediation is the proper, causal, mechanism regarding the relationship between the IV and DV. An excellent review of the challenges of establishing a robust mediation model is provided by Kline [2015], where he suggests the following as the minimally required elements of a mediation design:

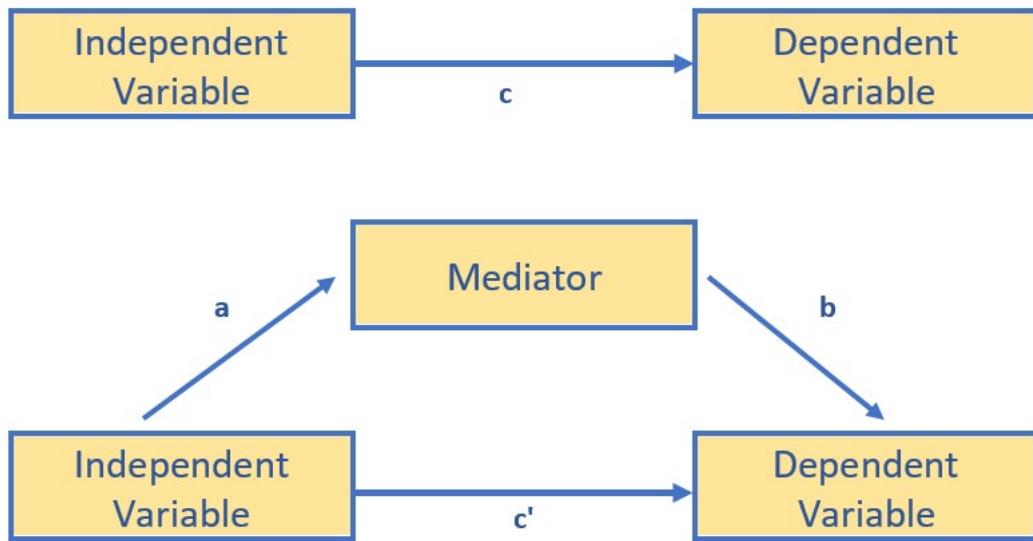


Figure 5.6: Image of conditional process analysis model where the mediator is hypothesized to change the a path; the path between the IV and mediator

- the IV is an experimental variable with random assignment to conditions;
- the mediator is an individual difference variable that is not manipulated and is measured at a later time; and
- the DV is measured at a third occasion

These criteria are in addition to the rather standard criteria for establishing causality [see [Stone-Romero and Rosopa, 2010](#), for a review]:

- temporal precedence,
- statistical covariation, and
- ruling out plausible rival hypotheses.

Some journals take this very seriously. In fact [FAQs](#) in the Journal of Vocational Behavior make it clear that they will very rarely publish a “mediation manuscript” unless it has a minimum of three waves.

Working through a mediation will help operationalize these concepts.

## 5.4 Super Simple Mediation in *lavaan*: A focus on the mechanics

The lavaan tutorial [[Rosseel, 2020](#)] provides a helpful model of how writing code to estimate an indirect effect. Using the lavaan tutorial as our guide, let’s start with just a set of fake data with variable names that represent X (predictor, IV, antecedent), M (mediator, atencedent, consequent), and Y (outcome, DV, consequent).

### 5.4.1 Simulate Fake Data

The code below is asking to create a dataset with a sample size of 100. The dataset has 3 variables, conveniently named X (predictor, antecedent, IV), M (mediator), and Y (outcome, consequent, DV). The R code asks for random selection of numbers with a normal distribution. You can see that the M variable will be related to the X variable by + .5; and the Y variable will be related to the M variable by + .7. This rather ensures a statistically significant indirect effect.

```
set.seed(230916)
X <- rnorm(100)
M <- 0.5 * X + rnorm(100)
Y <- 0.7 * M + rnorm(100)
Data <- data.frame(X = X, Y = Y, M = M)
```

### 5.4.2 Specify Mediation Model

The package we are using is *lavaan*. Hayes' model is *path analysis*, which can be a form of structural equation modeling. As a quick reminder, in SPSS, PROCESS is limited to ordinary least squares regression. We will use maximum likelihood estimators for the Hayes/PROCESS examples, but *lavaan* can take us further than PROCESS because

- We can (and, in later chapters, will) do latent variable modeling.
- We can have more specificity and flexibility than the prescribed PROCESS models allow. I say this with all due respect to Hayes – there is also a good deal of flexibility to be able to add multiple mediators and covariates within most of the Hayes' prescribed models.

Hayes text is still a great place to start because the conceptual and procedural information is clear and transferable to the R environment.

Our atheoretical dataset makes it easy to identify which variable belongs in each role (X,Y,M). When specifying the paths in lavaan, here's what to keep in mind:

- Name your model/object (below is X, “<-” means “is defined by”)
- The model exists between 2 single quotation marks (the odd looking ' and ' at the beginning and end).
- The # of regression equations you need depends on the # of variables that have arrows pointing to them. In a simple mediation, there are 3 variables with 2 variables having arrows pointing to them – need 2 regression equations:
  - one for the Mediator
  - one for the DV (Y)
- Operator for a regression analysis is the (tilde, ~)
- DV goes on left
  - In first equation we regress both the X and M onto Y
  - In second equation we regress M onto X

- The asterisk (\*) is a handy tool to label variables (don't confuse it as defining an interaction); this labeling as a, b, and c\_p (in traditional mediation, the total effect is labeled with a and the direct effect is c'[c prime], but the script won't allow an extra single quotation mark, hence c\_p) is super helpful in interpreting the output
- The indirect effect is created by multiplying the a and b paths.
- The “:=” sign is used when creating a new variable that is a function of variables in the model, but not in the dataset (i.e., the a and b path).

After specifying the model, we create an object that holds our results from the SEM. To obtain all the results from our of indirect effects, we also need to print a summary of the fit statistics, standardized estimates, r-squared, and confidence intervals.

*Other authors will write the model code more sensibly, predicting the mediator first, and then the Y variable. However, I found that by doing it this way, the semPlot produces a more sensible figure.*

Also, because we set a random seed, you should get the same results, but if it differs a little, don't panic. Also, in Hayes text the direct path from X to Y is c' (“c prime”; where as c is reserved for the total effect of X on Y).

Let's run the whole model.

```
model <- "
  Y ~ b*M + c_p*X
  M ~ a*X

  indirect := a*b
  direct   := c_p
  total_c  := c_p + (a*b)
  "

fit <- lavaan::sem(model, data = Data, se = "bootstrap", missing = "fiml")
FDsummary <- lavaan::summary(fit, standardized = T, rsq = T, fit = TRUE,
  ci = TRUE)
FD_ParamEsts <- lavaan::parameterEstimates(fit, boot.ci.type = "bca.simple",
  standardized = TRUE)
FDsummary

## lavaan 0.6.16 ended normally after 1 iteration
##
##   Estimator                               ML
##   Optimization method                     NLMINB
##   Number of model parameters             7
## 
##   Number of observations                 100
##   Number of missing patterns              1
## 
##   Model Test User Model:
## 
##   Test statistic                           0.000
```

```

## Degrees of freedom 0
##
## Model Test Baseline Model:
##
## Test statistic 66.380
## Degrees of freedom 3
## P-value 0.000
##
## User Model versus Baseline Model:
##
## Comparative Fit Index (CFI) 1.000
## Tucker-Lewis Index (TLI) 1.000
##
## Robust Comparative Fit Index (CFI) 1.000
## Robust Tucker-Lewis Index (TLI) 1.000
##
## Loglikelihood and Information Criteria:
##
## Loglikelihood user model (H0) -279.032
## Loglikelihood unrestricted model (H1) -279.032
##
## Akaike (AIC) 572.064
## Bayesian (BIC) 590.301
## Sample-size adjusted Bayesian (SABIC) 568.193
##
## Root Mean Square Error of Approximation:
##
## RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: RMSEA <= 0.050 NA
## P-value H_0: RMSEA >= 0.080 NA
##
## Robust RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: Robust RMSEA <= 0.050 NA
## P-value H_0: Robust RMSEA >= 0.080 NA
##
## Standardized Root Mean Square Residual:
##
## SRMR 0.000
##
## Parameter Estimates:
##
## Standard errors Bootstrap
## Number of requested bootstrap draws 1000
## Number of successful bootstrap draws 1000

```

```

## 
## Regressions:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   Y ~
##     M          (b)    0.708    0.082   8.642   0.000    0.552    0.867
##     X          (c_p) -0.107    0.108  -0.986   0.324   -0.329    0.092
##   M ~
##     X          (a)    0.513    0.093   5.510   0.000    0.329    0.704
##   Std.lv  Std.all
## 
##     0.708    0.639
##    -0.107   -0.080
## 
##     0.513    0.426
## 
## Intercepts:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   .Y
##     .M    -0.022    0.099  -0.221   0.825   -0.205    0.181
##   Std.lv  Std.all
##    -0.022   -0.018
##    -0.031   -0.028
## 
## Variances:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   .Y    0.927    0.125   7.436   0.000    0.664    1.166
##   .M    0.981    0.124   7.924   0.000    0.724    1.224
##   Std.lv  Std.all
##    0.927    0.629
##    0.981    0.818
## 
## R-Square:
##                               Estimate
##   Y      0.371
##   M      0.182
## 
## Defined Parameters:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   indirect    0.363    0.078   4.661   0.000    0.226    0.531
##   direct     -0.107    0.108  -0.985   0.325   -0.329    0.092
##   total_c    0.257    0.118   2.169   0.030    0.030    0.476
##   Std.lv  Std.all
##    0.363    0.272
##   -0.107   -0.080
##    0.257    0.192

```

## FD\_ParamEsts

```

##      lhs op      rhs   label    est     se      z pvalue ci.lower ci.upper
## 1      Y ~       M      b  0.708  0.082  8.642  0.000  0.544   0.861
## 2      Y ~       X    c_p -0.107  0.108 -0.986  0.324 -0.314   0.109
## 3      M ~       X      a  0.513  0.093  5.510  0.000  0.322   0.697
## 4      Y ~~      Y      0.927  0.125  7.436  0.000  0.740   1.262
## 5      M ~~      M      0.981  0.124  7.924  0.000  0.772   1.276
## 6      X ~~      X      0.827  0.000     NA     NA  0.827   0.827
## 7      Y ~1      Y      -0.022  0.099 -0.221  0.825 -0.199   0.194
## 8      M ~1      M      -0.031  0.100 -0.310  0.756 -0.230   0.160
## 9      X ~1      X      -0.005  0.000     NA     NA -0.005  -0.005
## 10 indirect :=  a*b indirect  0.363  0.078  4.661  0.000  0.227   0.532
## 11 direct :=   c_p   direct -0.107  0.108 -0.985  0.325 -0.314   0.109
## 12 total_c := c_p+(a*b) total_c  0.257  0.118  2.169  0.030  0.031   0.478
##      std.lv std.all std.nox
## 1    0.708   0.639   0.639
## 2   -0.107  -0.080  -0.088
## 3    0.513   0.426   0.469
## 4    0.927   0.629   0.629
## 5    0.981   0.818   0.818
## 6    0.827   1.000   0.827
## 7   -0.022  -0.018  -0.018
## 8   -0.031  -0.028  -0.028
## 9   -0.005  -0.005  -0.005
## 10   0.363   0.272   0.299
## 11  -0.107  -0.080  -0.088
## 12   0.257   0.192   0.211

```

## 5.4.3 Interpret the Output

Note that in the script we ask (and get) two sets of parameter estimates. The second set (in the really nice dataframe) includes bootstrapped, bias-corrected confidence intervals. Bias-corrected confidence intervals have the advantage of being more powerful and bias-free. Note, though, that when the CI crosses 0, the effect is NS.

So let's look at this step-by-step.

- Overall, our model accounted for 37% of the variance in the IV and 18% of the variance in the mediator.
- a path =  $B = 0.513, p < 0.001$
- b path =  $0.708, p < 0.001$
- the indirect effect is a product of the a and b paths ( $0.513 * 0.708 = 0.363$ ); while we don't hand calculate it's significance, we see that it is  $p < 0.001$ .
- the direct effect (c', c prime, or c\_p) is the isolated effect of X on Y when including M as a predictor. We hope this value is *lower* than the total effect because this means that including

M shared some of the variance in predicting Y:  $c' = -0.107, p = 0.346$ , and it is no longer significant.

- we also see the total effect; this value is

- identical to the value of simply predicting Y on X (with no M it the model)
- the value of  $a(b) + c_p$ :  $(0.513 * 0.708) + (-0.107) = 0.257; (p = 0.035)$

Here's a demonstration that the total effect is, simply, predicting Y from X:

```
fitXY <- lm(Y ~ X, data = Data)
summary(fitXY)
```

```
##
## Call:
## lm(formula = Y ~ X, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36350 -0.90598 -0.07158  0.74879  2.52079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04374    0.12035  -0.363   0.7171
## X           0.25668    0.13237   1.939   0.0554 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.204 on 98 degrees of freedom
## Multiple R-squared:  0.03695,    Adjusted R-squared:  0.02712
## F-statistic:  3.76 on 1 and 98 DF,  p-value: 0.05537
```

In a simple model such as this, it is also the same value as the bivariate correlation. The only trick is that the bivariate correlation produces a standardized result; so it would be the  $\beta$ .

```
library(psych)
XY_r <- corr.test(Data[c("Y", "X")])
XY_r
```

```
## Call:corr.test(x = Data[c("Y", "X")])
## Correlation matrix
##      Y     X
## Y  1.00  0.19
## X  0.19  1.00
## Sample Size
## [1] 100
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##      Y     X
```

```
## Y 0.00 0.06
## X 0.06 0.00
##
## To see confidence intervals of the correlations, print with the short=FALSE option
```

#### 5.4.4 A Figure and Table

We can use the package [tidySEM](#) to create a figure that includes the values on the path.

Here's what the base package gets us

```
# only worked when I used the library to turn on all these pkgs
library(lavaan)
```

```
## This is lavaan 0.6-16
## lavaan is FREE software! Please report any bugs.
```

```
##
## Attaching package: 'lavaan'
```

```
## The following object is masked from 'package:psych':
##      cor2cov
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##      filter, lag
```

```
## The following objects are masked from 'package:base':
##      intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##      %+%, alpha
```

```
library(tidySEM)

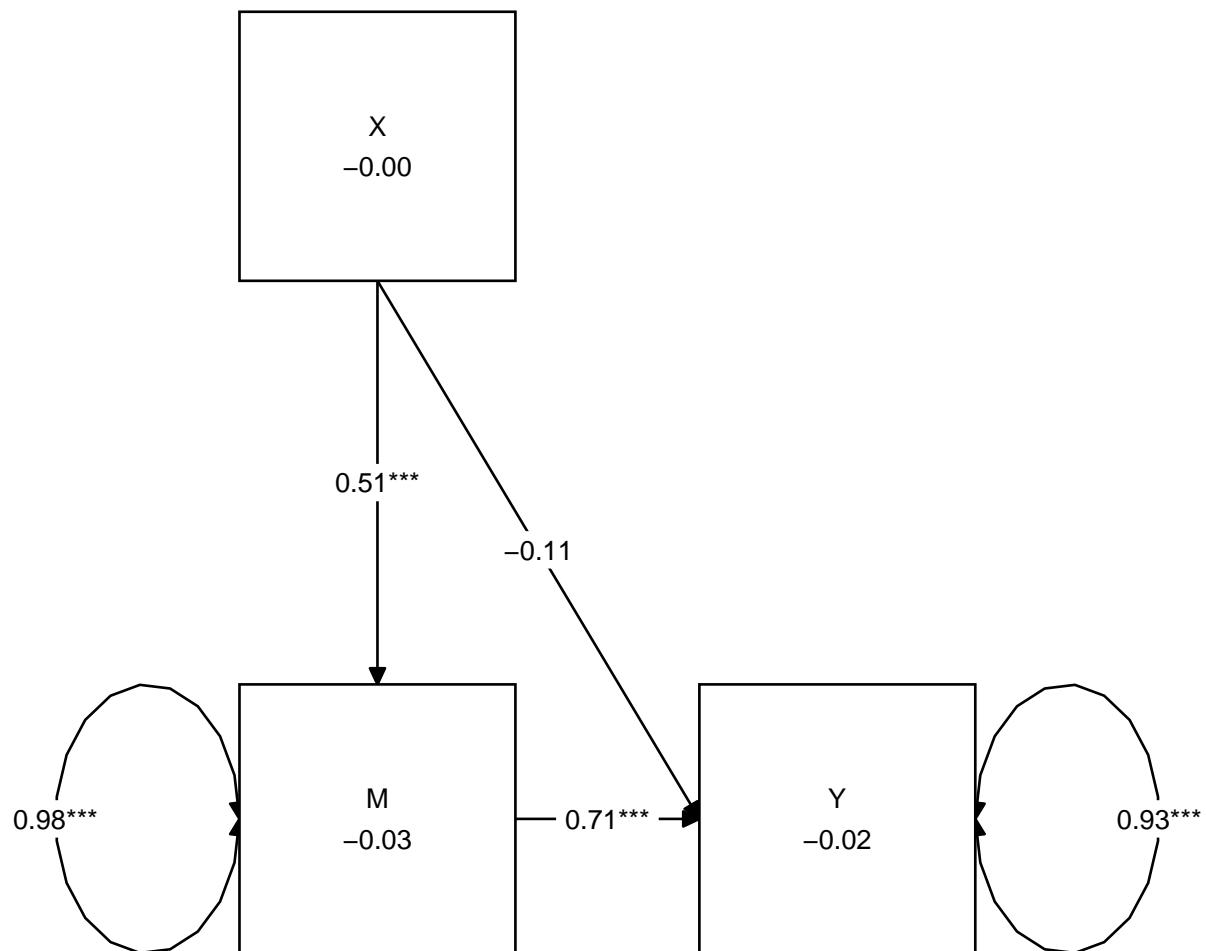
## Loading required package: OpenMx

##
## Attaching package: 'OpenMx'

## The following object is masked from 'package:psych':
## 
##     tr

## Registered S3 method overwritten by 'tidySEM':
##   method      from
##   predict.MxModel  OpenMx

tidySEM::graph_sem(model = fit)
```



Hayes has great examples of APA style tables that have become the standard way to communicate

results. I haven't yet found a package that will turn this output into a journal-ready table, however with a little tinkering, we can approximate one of the standard tables. This code lets us understand the label names and how they are mapped

```
tidySEM::get_layout(fit)

##      [,1] [,2]
## [1,] "Y"  "X"
## [2,] NA   "M"
## attr(),"class")
## [1] "layout_matrix" "matrix"      "array"
```

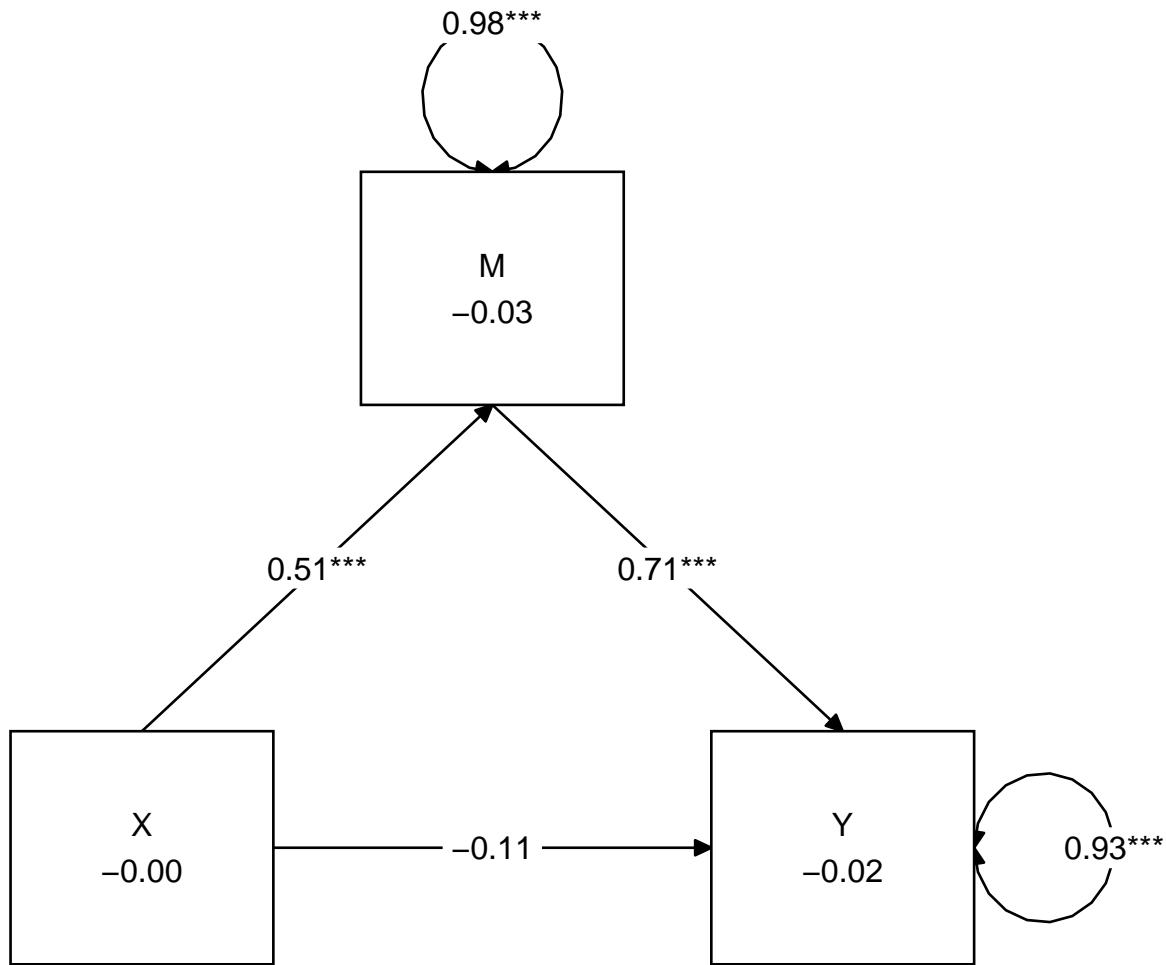
We can write code to remap them

```
med_map <- tidySEM::get_layout("", "M", "", "X", "", "Y", rows = 2)
med_map
```

```
##      [,1] [,2] [,3]
## [1,] ""   "M"  ""
## [2,] "X"  ""   "Y"
## attr(),"class")
## [1] "layout_matrix" "matrix"      "array"
```

We run again with our map and BOOM! Still needs tinkering for gorgeous, but hey!

```
tidySEM::graph_sem(fit, layout = med_map, rect_width = 1.5, rect_height = 1.25,
                   spacing_x = 2, spacing_y = 3, text_size = 4.5)
```



To assist in table preparation, it is possible to export the results to a .csv file that can be manipulated in Excel, Microsoft Word, or other program to prepare an APA style table.

```
write.csv(FD ParamEsts, file = "FakeDataOUT.csv")
```

Check with your discipline's journals to see how results of mediations are reported. Here's a version that I like.

Table 1

constant	$i_M$	0.031	0.098	0.753	$i_Y$	-0.022	0.099	0.826
Independent (X)	$a$	0.513	0.100	< 0.001	$c'$	-0.107	0.113	0.346
Mediator (M)					$b$	0.708	0.085	< 0.001
$R^2 = 18\%$				$R^2 = 37\%$				

Note. The value of the indirect effect was  $B = 0.363, SE = 0.084, p < 0.001, 95CI(0.226, 0.557)$

#### 5.4.5 Results

A simple mediation model examined the degree to which M mediated the relation of X on Y. Using the *lavaan* package (v 0.6-16) in R, coefficients for each path, the indirect effect, and total effects were calculated. These values are presented in Table 1 and illustrated in Figure 1. Results suggested that 18% of the variance in M and 37% of the variance in Y were accounted for in the model. The indirect effect ( $B = 0.363, SE = 0.084, p < 0.001$ ) was statistically significant; the direct effect ( $B = -0.107, SE = 0.113, p = 0.346$ ) was not. Comparing the nonsignificant direct effect to the statistically significant total effect ( $B = 0.257, SE = 0.121, p = 0.035$ ) is consistent with the notion that the effect of X on Y is explained through M.

### 5.5 Research Vignette

The research vignette comes from the Kim, Kendall, and Cheon's [2017], "Racial Microaggressions, Cultural Mistrust, and Mental Health Outcomes Among Asian American College Students." Participants were 156 Asian American undergraduate students in the Pacific Northwest. The researchers posited the a priori hypothesis that cultural mistrust would mediate the relationship between racial microaggressions and two sets of outcomes: mental health (e.g., depression, anxiety, well-being) and help-seeking.

Variables used in the study included:

- **REMS:** Racial and Ethnic Microaggressions Scale (Nadal, 2011). The scale includes 45 items on a 2-point scale where 0 indicates no experience of a microaggressive event and 1 indicates it was experienced at least once within the past six months. Higher scores indicate more experience of microaggressions.
- **CMI:** Cultural Mistrust Inventory (Terrell & Terrell, 1981). This scale was adapted to assess cultural mistrust harbored among Asian Americans toward individuals from the mainstream U.S. culture (e.g., Whites). The CMI includes 47 items on a 7-point scale where higher scores indicate a higher degree of cultural mistrust.
- **ANX, DEP, PWB:** Subscales of the Mental Health Inventory (Veit & Ware, 1983) that assess the mental health outcomes of anxiety (9 items), depression (4 items), and psychological well-being (14 items). Higher scores (on a 6 point scale) indicate stronger endorsement of the mental health outcome being assessed.

- **HlpSkg:** The Attitudes Toward Seeking Professional Psychological Help – Short Form (Fischer & Farina, 1995) includes 10 items on a 4-point scale (0 = disagree, 3 = agree) where higher scores indicate more favorable attitudes toward help seeking.

### 5.5.1 Data Simulation

We used the *lavaan::simulateData* function for the simulation. If you have taken psychometrics, you may recognize the code as one that creates latent variables from item-level data. In trying to be as authentic as possible, we retrieved factor loadings from psychometrically oriented articles that evaluated the measures [Nadal, 2011, Veit and Ware, 1983]. For all others we specified a factor loading of 0.80. We then approximated the *measurement model* by specifying the correlations between the latent variable. We sourced these from the correlation matrix from the research vignette [Kim et al., 2017]. The process created data with multiple decimals and values that exceeded the boundaries of the variables. For example, in all scales there were negative values. Therefore, the final element of the simulation was a linear transformation that rescaled the variables back to the range described in the journal article and rounding the values to integer (i.e., with no decimal places).

```
# Entering the intercorrelations, means, and standard deviations from
# the journal article
Kim_generating_model <- "
  ##measurement model
  REMS =~ .82*Inf32 + .75*Inf38 + .74*Inf21 + .72*Inf17 + .69*Inf9 + .61*Inf36 + .51*In
  CMI =~ .8*cmi1 + .8*cmi2 + .8*cmi3 + .8*cmi4 + .8*cmi5 + .8*cmi6 + .8*cmi7 + .8*cmi8 +
  ANX =~ .80*Anx1 + .80*Anx2 + .77*Anx3 + .74*Anx4 + .74*Anx5 + .69*Anx6 + .69*Anx7 + .
  DEP =~ .74*Dep1 + .83*Dep2 + .82*Dep3 + .74*Dep4
  PWB =~ .83*pwb1 + .72*pwb2 + .67*pwb3 + .79*pwb4 + .77*pwb5 + .75*pwb6 + .74*pwb7 + .7
  HlpSkg =~ .8*hlpstk1 + .8*hlpstk2 + .8*hlpstk3 + .8*hlpstk4 + .8*hlpstk5 + .8*hlpstk6

  # Means
  REMS ~ 0.34*1
  CMI ~ 3*1
  ANX ~ 2.98*1
  DEP ~ 2.36*1
  PWB ~ 3.5*1
  HlpSkg ~ 1.64*1
  # Correlations (ha!)
  REMS ~ 0.58*CMI
  REMS ~ 0.26*ANX
  REMS ~ 0.34*DEP
  REMS ~ -0.25*PWB
  REMS ~ -0.02*HlpSkg
  CMI ~ 0.12*ANX
  CMI ~ 0.19*DEP
```

```

CMI ~ -0.28*PWB
CMI ~ 0*HlpSkg
ANX ~ 0.66*DEP
ANX ~ -0.55*PWB
ANX ~ 0.07*HlpSkg
DEP ~ -0.66*PWB
DEP ~ 0.05*HlpSkg
PWB ~ 0.08*HlpSkg
""

set.seed(230916)
dfKim <- lavaan::simulateData(model = Kim_generating_model, model.type = "sem",
  meanstructure = T, sample.nobs = 156, standardized = FALSE)
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr    1.5.0
## v lubridate  1.9.2     v tibble     3.2.1
## v purrr     1.0.1     v tidyverse  1.3.0
## v readr      2.1.4
## -- Conflicts ----- tidyverse_conflicts() --
## x ggplot2::%+%( ) masks psych::%+%( )
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco

# Kim_df_latent <- Kim_df_latent %>% round(0) %>% abs()

dfKim$Inf32 <- scales::rescale(dfKim$Inf32, c(0, 1))
dfKim$Inf38 <- scales::rescale(dfKim$Inf38, c(0, 1))
dfKim$Inf21 <- scales::rescale(dfKim$Inf21, c(0, 1))
dfKim$Inf17 <- scales::rescale(dfKim$Inf17, c(0, 1))
dfKim$Inf9 <- scales::rescale(dfKim$Inf9, c(0, 1))
dfKim$Inf36 <- scales::rescale(dfKim$Inf36, c(0, 1))
dfKim$Inf5 <- scales::rescale(dfKim$Inf5, c(0, 1))
dfKim$Inf22 <- scales::rescale(dfKim$Inf22, c(0, 1))
dfKim$SClass6 <- scales::rescale(dfKim$SClass6, c(0, 1))
dfKim$SClass31 <- scales::rescale(dfKim$SClass31, c(0, 1))
dfKim$SClass8 <- scales::rescale(dfKim$SClass8, c(0, 1))
dfKim$SClass40 <- scales::rescale(dfKim$SClass40, c(0, 1))
dfKim$SClass2 <- scales::rescale(dfKim$SClass2, c(0, 1))
dfKim$SClass34 <- scales::rescale(dfKim$SClass34, c(0, 1))
dfKim$SClass11 <- scales::rescale(dfKim$SClass11, c(0, 1))
dfKim$mInv27 <- scales::rescale(dfKim$mInv27, c(0, 1))
dfKim$mInv30 <- scales::rescale(dfKim$mInv30, c(0, 1))
dfKim$mInv39 <- scales::rescale(dfKim$mInv39, c(0, 1))

```

```

dfKim$mInv7 <- scales::rescale(dfKim$mInv7, c(0, 1))
dfKim$mInv26 <- scales::rescale(dfKim$mInv26, c(0, 1))
dfKim$mInv33 <- scales::rescale(dfKim$mInv33, c(0, 1))
dfKim$mInv4 <- scales::rescale(dfKim$mInv4, c(0, 1))
dfKim$mInv14 <- scales::rescale(dfKim$mInv14, c(0, 1))
dfKim$mInv10 <- scales::rescale(dfKim$mInv10, c(0, 1))
dfKim$Exot3 <- scales::rescale(dfKim$Exot3, c(0, 1))
dfKim$Exot29 <- scales::rescale(dfKim$Exot29, c(0, 1))
dfKim$Exot45 <- scales::rescale(dfKim$Exot45, c(0, 1))
dfKim$Exot35 <- scales::rescale(dfKim$Exot35, c(0, 1))
dfKim$Exot42 <- scales::rescale(dfKim$Exot42, c(0, 1))
dfKim$Exot23 <- scales::rescale(dfKim$Exot23, c(0, 1))
dfKim$Exot13 <- scales::rescale(dfKim$Exot13, c(0, 1))
dfKim$Exot20 <- scales::rescale(dfKim$Exot20, c(0, 1))
dfKim$Exot43 <- scales::rescale(dfKim$Exot43, c(0, 1))
dfKim$mEnv37 <- scales::rescale(dfKim$mEnv37, c(0, 1))
dfKim$mEnv24 <- scales::rescale(dfKim$mEnv24, c(0, 1))
dfKim$mEnv19 <- scales::rescale(dfKim$mEnv19, c(0, 1))
dfKim$mEnv28 <- scales::rescale(dfKim$mEnv28, c(0, 1))
dfKim$mEnv18 <- scales::rescale(dfKim$mEnv18, c(0, 1))
dfKim$mEnv41 <- scales::rescale(dfKim$mEnv41, c(0, 1))
dfKim$mEnv12 <- scales::rescale(dfKim$mEnv12, c(0, 1))
dfKim$mWork25 <- scales::rescale(dfKim$mWork25, c(0, 1))
dfKim$mWork15 <- scales::rescale(dfKim$mWork15, c(0, 1))
dfKim$mWork1 <- scales::rescale(dfKim$mWork1, c(0, 1))
dfKim$mWork16 <- scales::rescale(dfKim$mWork16, c(0, 1))
dfKim$mWork44 <- scales::rescale(dfKim$mWork44, c(0, 1))

dfKim$cmi1 <- scales::rescale(dfKim$cmi1, c(1, 7))
dfKim$cmi2 <- scales::rescale(dfKim$cmi2, c(1, 7))
dfKim$cmi3 <- scales::rescale(dfKim$cmi3, c(1, 7))
dfKim$cmi4 <- scales::rescale(dfKim$cmi4, c(1, 7))
dfKim$cmi5 <- scales::rescale(dfKim$cmi5, c(1, 7))
dfKim$cmi6 <- scales::rescale(dfKim$cmi6, c(1, 7))
dfKim$cmi7 <- scales::rescale(dfKim$cmi7, c(1, 7))
dfKim$cmi8 <- scales::rescale(dfKim$cmi8, c(1, 7))
dfKim$cmi9 <- scales::rescale(dfKim$cmi9, c(1, 7))
dfKim$cmi10 <- scales::rescale(dfKim$cmi10, c(1, 7))
dfKim$cmi11 <- scales::rescale(dfKim$cmi11, c(1, 7))
dfKim$cmi12 <- scales::rescale(dfKim$cmi12, c(1, 7))
dfKim$cmi13 <- scales::rescale(dfKim$cmi13, c(1, 7))
dfKim$cmi14 <- scales::rescale(dfKim$cmi14, c(1, 7))
dfKim$cmi15 <- scales::rescale(dfKim$cmi15, c(1, 7))
dfKim$cmi16 <- scales::rescale(dfKim$cmi16, c(1, 7))
dfKim$cmi17 <- scales::rescale(dfKim$cmi17, c(1, 7))
dfKim$cmi18 <- scales::rescale(dfKim$cmi18, c(1, 7))
dfKim$cmi19 <- scales::rescale(dfKim$cmi19, c(1, 7))

```

```
dfKim$cmi20 <- scales::rescale(dfKim$cmi20, c(1, 7))
dfKim$cmi21 <- scales::rescale(dfKim$cmi21, c(1, 7))
dfKim$cmi22 <- scales::rescale(dfKim$cmi22, c(1, 7))
dfKim$cmi23 <- scales::rescale(dfKim$cmi23, c(1, 7))
dfKim$cmi24 <- scales::rescale(dfKim$cmi24, c(1, 7))
dfKim$cmi25 <- scales::rescale(dfKim$cmi25, c(1, 7))
dfKim$cmi26 <- scales::rescale(dfKim$cmi26, c(1, 7))
dfKim$cmi27 <- scales::rescale(dfKim$cmi27, c(1, 7))
dfKim$cmi28 <- scales::rescale(dfKim$cmi28, c(1, 7))
dfKim$cmi29 <- scales::rescale(dfKim$cmi29, c(1, 7))
dfKim$cmi30 <- scales::rescale(dfKim$cmi30, c(1, 7))
dfKim$cmi31 <- scales::rescale(dfKim$cmi31, c(1, 7))
dfKim$cmi32 <- scales::rescale(dfKim$cmi32, c(1, 7))
dfKim$cmi33 <- scales::rescale(dfKim$cmi33, c(1, 7))
dfKim$cmi34 <- scales::rescale(dfKim$cmi34, c(1, 7))
dfKim$cmi35 <- scales::rescale(dfKim$cmi35, c(1, 7))
dfKim$cmi36 <- scales::rescale(dfKim$cmi36, c(1, 7))
dfKim$cmi37 <- scales::rescale(dfKim$cmi37, c(1, 7))
dfKim$cmi38 <- scales::rescale(dfKim$cmi38, c(1, 7))
dfKim$cmi39 <- scales::rescale(dfKim$cmi39, c(1, 7))
dfKim$cmi40 <- scales::rescale(dfKim$cmi40, c(1, 7))
dfKim$cmi41 <- scales::rescale(dfKim$cmi41, c(1, 7))
dfKim$cmi42 <- scales::rescale(dfKim$cmi42, c(1, 7))
dfKim$cmi43 <- scales::rescale(dfKim$cmi43, c(1, 7))
dfKim$cmi44 <- scales::rescale(dfKim$cmi44, c(1, 7))
dfKim$cmi45 <- scales::rescale(dfKim$cmi45, c(1, 7))
dfKim$cmi46 <- scales::rescale(dfKim$cmi46, c(1, 7))
dfKim$cmi47 <- scales::rescale(dfKim$cmi47, c(1, 7))

dfKim$Anx1 <- scales::rescale(dfKim$Anx1, c(1, 5))
dfKim$Anx2 <- scales::rescale(dfKim$Anx2, c(1, 5))
dfKim$Anx3 <- scales::rescale(dfKim$Anx3, c(1, 5))
dfKim$Anx4 <- scales::rescale(dfKim$Anx4, c(1, 5))
dfKim$Anx5 <- scales::rescale(dfKim$Anx5, c(1, 5))
dfKim$Anx6 <- scales::rescale(dfKim$Anx6, c(1, 5))
dfKim$Anx7 <- scales::rescale(dfKim$Anx7, c(1, 5))
dfKim$Anx8 <- scales::rescale(dfKim$Anx8, c(1, 5))
dfKim$Anx9 <- scales::rescale(dfKim$Anx9, c(1, 5))

dfKim$Dep1 <- scales::rescale(dfKim$Dep1, c(1, 5))
dfKim$Dep2 <- scales::rescale(dfKim$Dep2, c(1, 5))
dfKim$Dep3 <- scales::rescale(dfKim$Dep3, c(1, 5))
dfKim$Dep4 <- scales::rescale(dfKim$Dep4, c(1, 5))

dfKim$pwb1 <- scales::rescale(dfKim$pwb1, c(1, 5))
dfKim$pwb2 <- scales::rescale(dfKim$pwb2, c(1, 5))
dfKim$pwb3 <- scales::rescale(dfKim$pwb3, c(1, 5))
```

```

dfKim$pwb4 <- scales::rescale(dfKim$pwb4, c(1, 5))
dfKim$pwb5 <- scales::rescale(dfKim$pwb5, c(1, 5))
dfKim$pwb6 <- scales::rescale(dfKim$pwb6, c(1, 5))
dfKim$pwb7 <- scales::rescale(dfKim$pwb7, c(1, 5))
dfKim$pwb8 <- scales::rescale(dfKim$pwb8, c(1, 5))
dfKim$pwb9 <- scales::rescale(dfKim$pwb9, c(1, 5))
dfKim$pwb10 <- scales::rescale(dfKim$pwb10, c(1, 5))
dfKim$pwb11 <- scales::rescale(dfKim$pwb11, c(1, 5))

dfKim$hlpstk1 <- scales::rescale(dfKim$hlpstk1, c(0, 3))
dfKim$hlpstk2 <- scales::rescale(dfKim$hlpstk2, c(0, 3))
dfKim$hlpstk3 <- scales::rescale(dfKim$hlpstk3, c(0, 3))
dfKim$hlpstk4 <- scales::rescale(dfKim$hlpstk4, c(0, 3))
dfKim$hlpstk5 <- scales::rescale(dfKim$hlpstk5, c(0, 3))
dfKim$hlpstk6 <- scales::rescale(dfKim$hlpstk6, c(0, 3))
dfKim$hlpstk7 <- scales::rescale(dfKim$hlpstk7, c(0, 3))
dfKim$hlpstk8 <- scales::rescale(dfKim$hlpstk8, c(0, 3))
dfKim$hlpstk9 <- scales::rescale(dfKim$hlpstk9, c(0, 3))
dfKim$hlpstk10 <- scales::rescale(dfKim$hlpstk10, c(0, 3))

# psych::describe(dfKim)

library(tidyverse)
dfKim <- dfKim %>%
  round(0)

# I tested the rescaling the correlation between original and
# rescaled variables is 1.0 Kim_df_latent$INF32 <-
# scales::rescale(Kim_df_latent$Inf32, c(0, 1))
# cor.test(Kim_df_latent$Inf32, Kim_df_latent$INF32,
# method='pearson')

# Checking our work against the original correlation matrix
# round(cor(Kim_df),3)

```

The script below allows you to store the simulated data as a file on your computer. This is optional – the entire lesson can be worked with the simulated data.

If you prefer the .rds format, use this script (remove the hashtags). The .rds format has the advantage of preserving any formatting of variables. A disadvantage is that you cannot open these files outside of the R environment.

Script to save the data to your computer as an .rds file.

```
#saveRDS(dfKim, 'dfKim.rds')
```

Once saved, you could clean your environment and bring the data back in from its .csv format.

```
# dfKim<- readRDS('dfKim.rds')
```

If you prefer the .csv format (think “Excel lite”) use this script (remove the hashtags). An advantage of the .csv format is that you can open the data outside of the R environment. A disadvantage is that it may not retain any formatting of variables

Script to save the data to your computer as a .csv file.

```
#write.table(dfKim, file = 'dfKim.csv', sep = ',', col.names=TRUE, row.names=FALSE)
```

Once saved, you could clean your environment and bring the data back in from its .csv format.

```
# dfKim<- read.csv ('dfKim.csv', header = TRUE)
```

### 5.5.2 Scrubbing, Scoring, and Data Diagnostics

Because the focus of this lesson is on simple mediation, we have used simulated data. If this were real, raw, data, it would be important to [scrub](#), [score](#), and conduct [data diagnostics](#) to evaluate the suitability of the data for the proposes analyses.

Because we are working with item level data we first need to score the scales used in the researcher’s model/. Because we are using simulated data and the authors already reverse coded any items requiring recoding, we can omit that step.

As described in the [Scoring](#) chapter, we can calculate mean scores of these variables by first creating concatenated lists of variable names. Next we apply the `sjstats::mean_n` function to obtain mean scores when a given percentage (we’ll specify 80%) of variables are non-missing. We simulated a set of data that does not have missingness, none-the-less, this specification is useful in real-world settings.

```
PWB_vars <- c("pwb1", "pwb2", "pwb3", "pwb4", "pwb5", "pwb6", "pwb7", "pwb8",
  "pwb9", "pwb10")
ANX_vars <- c("Anx1", "Anx2", "Anx3", "Anx4", "Anx5", "Anx6", "Anx7", "Anx8",
  "Anx9")
CMI_vars <- c("cmi1", "cmi2", "cmi3", "cmi4", "cmi5", "cmi6", "cmi7", "cmi8",
  "cmi9", "cmi10", "cmi11", "cmi12", "cmi13", "cmi14", "cmi15", "cmi16",
  "cmi17", "cmi18", "cmi19", "cmi20", "cmi21", "cmi22", "cmi23", "cmi24",
  "cmi25", "cmi26", "cmi27", "cmi28", "cmi29", "cmi30", "cmi31", "cmi32",
  "cmi33", "cmi34", "cmi35", "cmi36", "cmi37", "cmi38", "cmi39", "cmi40",
  "cmi41", "cmi42", "cmi43", "cmi44", "cmi45", "cmi46", "cmi47")
REMS_vars <- c("Inf32", "Inf38", "Inf21", "Inf17", "Inf9", "Inf36", "Inf5",
  "Inf22", "SClass6", "SClass31", "SClass8", "SClass40", "SClass2", "SClass34",
  "SClass11", "mInv27", "mInv30", "mInv39", "mInv7", "mInv26", "mInv33",
  "mInv4", "mInv14", "mInv10", "Exot3", "Exot29", "Exot45", "Exot35",
  "Exot42", "Exot23", "Exot13", "Exot20", "Exot43", "mEnv37", "mEnv24",
  "mEnv19", "mEnv28", "mEnv18", "mEnv41", "mEnv12", "mWork25", "mWork15",
  "mWork1", "mWork16", "mWork44")
```

```
dfKim$PWB <- sjstats::mean_n(dfKim[, PWB_vars], 0.8)
dfKim$ANX <- sjstats::mean_n(dfKim[, ANX_vars], 0.8)
dfKim$CMI <- sjstats::mean_n(dfKim[, CMI_vars], 0.8)
dfKim$REMS <- sjstats::mean_n(dfKim[, REMS_vars], 0.8)
```

Now that we have scored our data, let's trim the variables to just those we need.

```
dfModel <- dplyr::select(dfKim, PWB, ANX, CMI, REMS)
```

Let's check a table of means, standards, and correlations to see if they align with the published article.

```
DescriptivesTable <- apaTables::apa.cor.table(dfModel, table.number = 1,
                                             show.sig.stars = TRUE, landscape = TRUE, filename = NA)
print(DescriptivesTable)
```

```
##
##
## Table 1
##
## Means, standard deviations, and correlations with confidence intervals
##
##
##      Variable M      SD      1          2          3
##      1. PWB   3.09  0.45
##
##      2. ANX   2.82  0.57 -.50**    [-.61, -.37]
##
##      3. CMI   3.94  0.77 -.49**    .43**    [-.60, -.36] [.30, .55]
##
##      4. REMS  0.51  0.29 -.47**    .58**    .58**    [-.59, -.34] [.47, .68]  [.47, .68]
##
##
## Note. M and SD are used to represent mean and standard deviation, respectively.
## Values in square brackets indicate the 95% confidence interval.
## The confidence interval is a plausible range of population correlations
## that could have caused the sample correlation (Cumming, 2014).
## * indicates p < .05. ** indicates p < .01.
##
```

While the patterns are similar, we can see some differences. This means that our simulated results are likely to have some difference than the results in the published article.

Comparison	Article	Simulation
PWB mean	3.50	3.93
ANX mean	2.98	2.82
CMI mean	3.00	3.94
REM mean	.34	.51
PWB ~ ANX	-0.55***	-0.50**
PWB ~ CMI	-0.28***	-0.49**
PWB ~ REMS	-0.25**	-0.47**
ANX ~ CMI	0.12	0.43**
ANX ~ REMS	0.26**	0.58**
CMI ~ REMS	0.59***	0.58**

There are a number of reasons I love the Kim et al. [2017] manuscript. One is that their approach was openly one that tested *alternate models*. Byrne [2016] credits Joreskog [Joreskog, 1993] with classifying the researcher's model testing approach in three ways. If a researcher uses a *strictly confirmatory* approach, they only test the proposed model and then accept or reject it without further alteration. While this is the tradition of null hypothesis significance testing (NHST), it contributes to the “file drawer problem” of unpublished, non-significant, findings. Additionally, the data are them discarded – potentially losing valuable resource. The *alternative models* approach is to propose a handful of competing models before beginning the analysis and then evaluating to see if one model is superior to the other. The third option is *model generating*. In this case the researcher begins with a theoretically proposed model. In the presence of poor fit, the researcher seeks to identify the source of misfit – respecifying it to best represent the sample data. The researcher must use caution to produce a model that fits well and is meaningful.

Several of the Kim et al. [2017] models were non-significant. To demonstrate a model that is statistically significant, I will test the hypothesis that racial microaggressions (REMS, the X variable) influence depression (DEP, the Y variable) through cultural mistrust (CMI, the M variable).

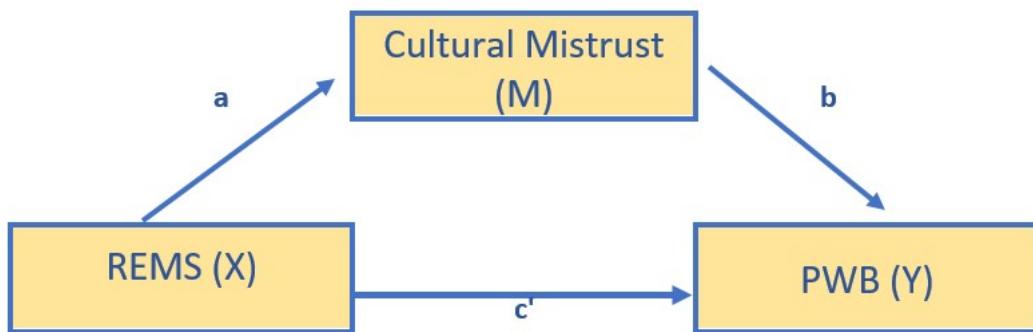


Figure 5.7: Image of the simple mediation model from Kim et al.

### 5.5.3 Specify the Model in *lavaan*

I am a big fan of “copying the model.” That is, I find *code that works* as a starting point. In specifying my model I used the simple mediation template from above. I

- replaced the Y, X, and M with variables names
- replacing the name of the df
- updated the object names (so I could use them in the same .rmd file)

```

modKim <- "
    PWB ~ b*CMI + c_p*REMS
    CMI ~a*REMS

    indirect := a*b
    direct   := c_p
    total_c  := c_p + (a*b)
    "

```

```

Kim_fit <- lavaan::sem(modKim, data = dfModel, se = "bootstrap", missing = "fiml")

Kim_summary <- summary(Kim_fit, standardized = T, rsq = T, fit = TRUE,
    ci = TRUE)
Kim_ParamEsts <- parameterEstimates(Kim_fit, boot.ci.type = "bca.simple",
    standardized = TRUE)
Kim_summary

## lavaan 0.6.16 ended normally after 1 iteration
##
##      Estimator                      ML
## Optimization method                NLMINB
## Number of model parameters          7
## 
##      Number of observations           156
##      Number of missing patterns        1
## 
## Model Test User Model:
## 
##      Test statistic                  0.000
##      Degrees of freedom                   0
## 
## Model Test Baseline Model:
## 
##      Test statistic                 119.320
##      Degrees of freedom                     3
##      P-value                           0.000
## 
## User Model versus Baseline Model:
## 
##      Comparative Fit Index (CFI)       1.000
##      Tucker-Lewis Index (TLI)         1.000
## 
##      Robust Comparative Fit Index (CFI) 1.000

```

```

## Robust Tucker-Lewis Index (TLI) 1.000
##
## Loglikelihood and Information Criteria:
##
## Loglikelihood user model (H0) -218.515
## Loglikelihood unrestricted model (H1) -218.515
##
## Akaike (AIC) 451.030
## Bayesian (BIC) 472.379
## Sample-size adjusted Bayesian (SABIC) 450.222
##
## Root Mean Square Error of Approximation:
##
## RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: RMSEA <= 0.050 NA
## P-value H_0: RMSEA >= 0.080 NA
##
## Robust RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: Robust RMSEA <= 0.050 NA
## P-value H_0: Robust RMSEA >= 0.080 NA
##
## Standardized Root Mean Square Residual:
##
## SRMR 0.000
##
## Parameter Estimates:
##
## Standard errors Bootstrap
## Number of requested bootstrap draws 1000
## Number of successful bootstrap draws 1000
##
## Regressions:
## Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## PWB ~
## CMI (b) -0.189 0.050 -3.809 0.000 -0.290 -0.087
## REMS (c_p) -0.453 0.138 -3.289 0.001 -0.733 -0.180
## CMI ~
## REMS (a) 1.576 0.169 9.300 0.000 1.253 1.930
## Std.lv Std.all
##
## -0.189 -0.323
## -0.453 -0.286
##
## 1.576 0.584

```

```

##  

## Intercepts:  

##  

##             Estimate Std. Err. z-value P(>|z|) ci.lower ci.upper  

## .PWB        4.066   0.173 23.488  0.000   3.696   4.385  

## .CMI        3.141   0.102 30.711  0.000   2.943   3.341  

## Std.lv Std.all  

## 4.066    9.004  

## 3.141    4.072  

##  

## Variances:  

##  

##             Estimate Std. Err. z-value P(>|z|) ci.lower ci.upper  

## .PWB        0.144   0.017  8.591  0.000   0.109   0.176  

## .CMI        0.392   0.041  9.659  0.000   0.311   0.473  

## Std.lv Std.all  

## 0.144    0.706  

## 0.392    0.659  

##  

## R-Square:  

##  

##             Estimate  

## PWB        0.294  

## CMI        0.341  

##  

## Defined Parameters:  

##  

##             Estimate Std. Err. z-value P(>|z|) ci.lower ci.upper  

## indirect     -0.298   0.087 -3.432  0.001  -0.479  -0.127  

## direct       -0.453   0.138 -3.288  0.001  -0.733  -0.180  

## total_c      -0.750   0.116 -6.490  0.000  -0.977  -0.517  

## Std.lv Std.all  

## -0.298   -0.188  

## -0.453   -0.286  

## -0.750   -0.475

```

### Kim\_ParamEsts

	lhs	op	rhs	label	est	se	z	pvalue	ci.lower	ci.upper
## 1	PWB	~	CMI	b	-0.189	0.050	-3.809	0.000	-0.288	-0.087
## 2	PWB	~	REMS	c_p	-0.453	0.138	-3.289	0.001	-0.751	-0.205
## 3	CMI	~	REMS	a	1.576	0.169	9.300	0.000	1.252	1.926
## 4	PWB	~~	PWB		0.144	0.017	8.591	0.000	0.114	0.182
## 5	CMI	~~	CMI		0.392	0.041	9.659	0.000	0.326	0.487
## 6	REMS	~~	REMS		0.082	0.000	NA	NA	0.082	0.082
## 7	PWB	~1			4.066	0.173	23.488	0.000	3.685	4.380
## 8	CMI	~1			3.141	0.102	30.711	0.000	2.952	3.353
## 9	REMS	~1			0.507	0.000	NA	NA	0.507	0.507
## 10	indirect	:=	a*b	indirect	-0.298	0.087	-3.432	0.001	-0.479	-0.127
## 11	direct	:=	c_p	direct	-0.453	0.138	-3.288	0.001	-0.751	-0.205
## 12	total_c	:=	c_p+(a*b)	total_c	-0.750	0.116	-6.490	0.000	-0.977	-0.517

```
##   std.lv std.all std.nox
## 1 -0.189 -0.323 -0.323
## 2 -0.453 -0.286 -1.002
## 3  1.576  0.584  2.043
## 4  0.144  0.706  0.706
## 5  0.392  0.659  0.659
## 6  0.082  1.000  0.082
## 7  4.066  9.004  9.004
## 8  3.141  4.072  4.072
## 9  0.507  1.775  0.507
## 10 -0.298 -0.188 -0.659
## 11 -0.453 -0.286 -1.002
## 12 -0.750 -0.475 -1.662
```

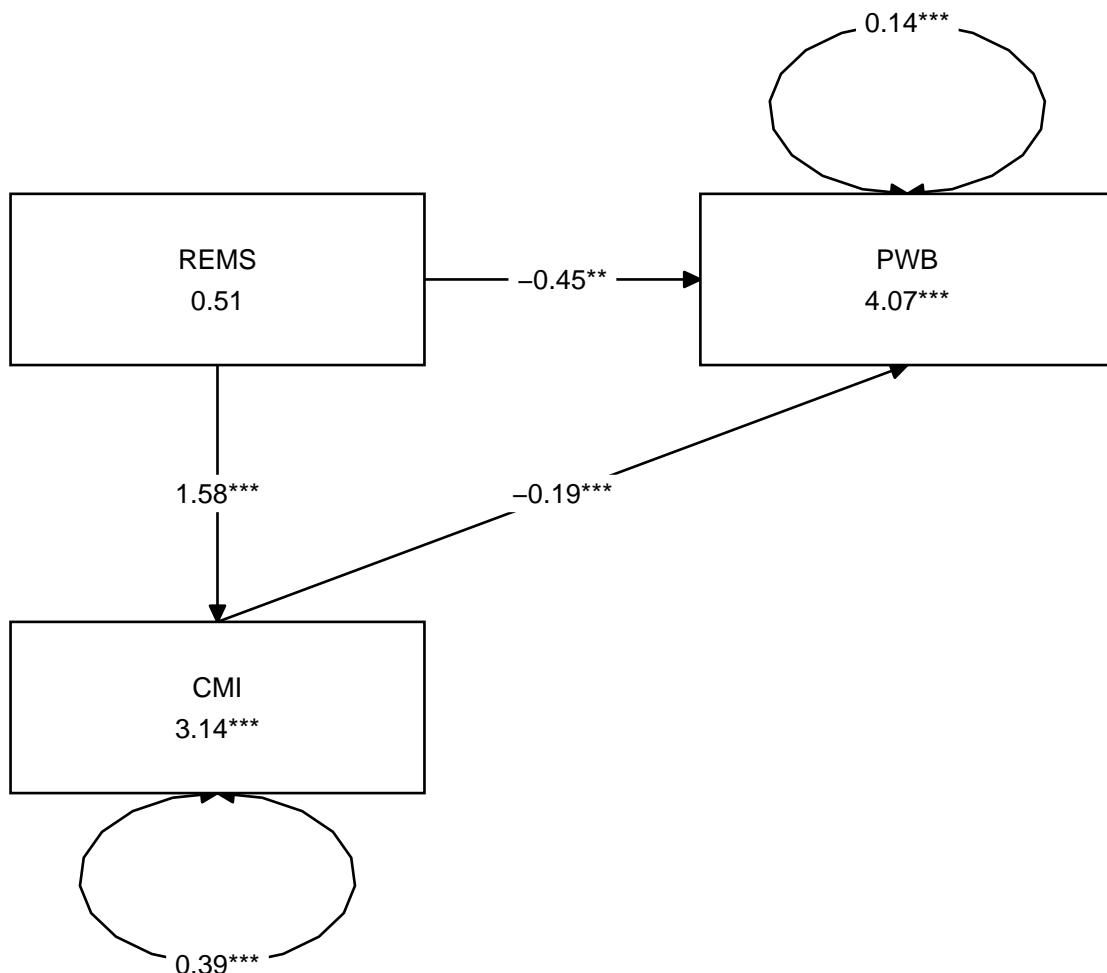
#### 5.5.4 Interpret the Output

- Overall, our model accounted for 29% of the variance in the independent variable, well-being, and 34% of the variance in the mediator, cultural mistrust.
- a path:  $B = 1.576, p < 0.001$
- b path:  $B = -0.189, p < 0.001$
- the indirect effect is a product of the a and b paths:  $B = -0.298, p = 0.001$ .
- The bias-corrected bootstrapped confidence intervals can sometimes be more lenient than  $p$  values; it is important they don't cross zero ( $95CI = -0.495, -0.136$ ). If 0.00 is included in the confidence interval, then we cannot be confident that the estimate is not, itself, zero.
- the direct effect ( $c'$ , c prime, or  $c_p$ ) is the isolated effect of X on Y when including M. We hope this value is lower than the total effect because it would mean that including M shared some of the variance in predicting Y. In our case the value for  $c'$  is:  $B = -0.453, p = 0.001$ . Unfortunately, they are significant and they are not markedly different from the total effect ( $B = -0.750, p < 0.001$ ).
- As a reminder, the total effect is
- identical to the value of simply predicting Y on X (with no M it the model)
- the value of  $a(b) + c_p$ :  $(1.576 * -0.189) + (-0.453) = -0.750; p < 0.001$

#### 5.5.5 A Figure and a Table

I make it a practice to immediately plot what I did. Because the plotting packages use our models, this can be a helpful self-check of our work.

```
# only worked when I used the library to turn on all these pkgs
library(lavaan)
library(dplyr)
library(ggplot2)
library(tidySEM)
tidySEM::graph_sem(model = Kim_fit)
```



Hayes has great examples of APA style tables that have become the standard way to communicate results. I haven't yet found a package that will turn this output into a journal-ready table, however with a little tinkering, we can approximate one of the standard tables. This code lets us understand the label names and how they are mapped

```
tidySEM::get_layout(Kim_fit)
```

```
##      [,1]  [,2]  [,3]
## [1,] "PWB" "CMI" "REMS"
## attr(),"class")
## [1] "layout_matrix" "matrix"       "array"
```

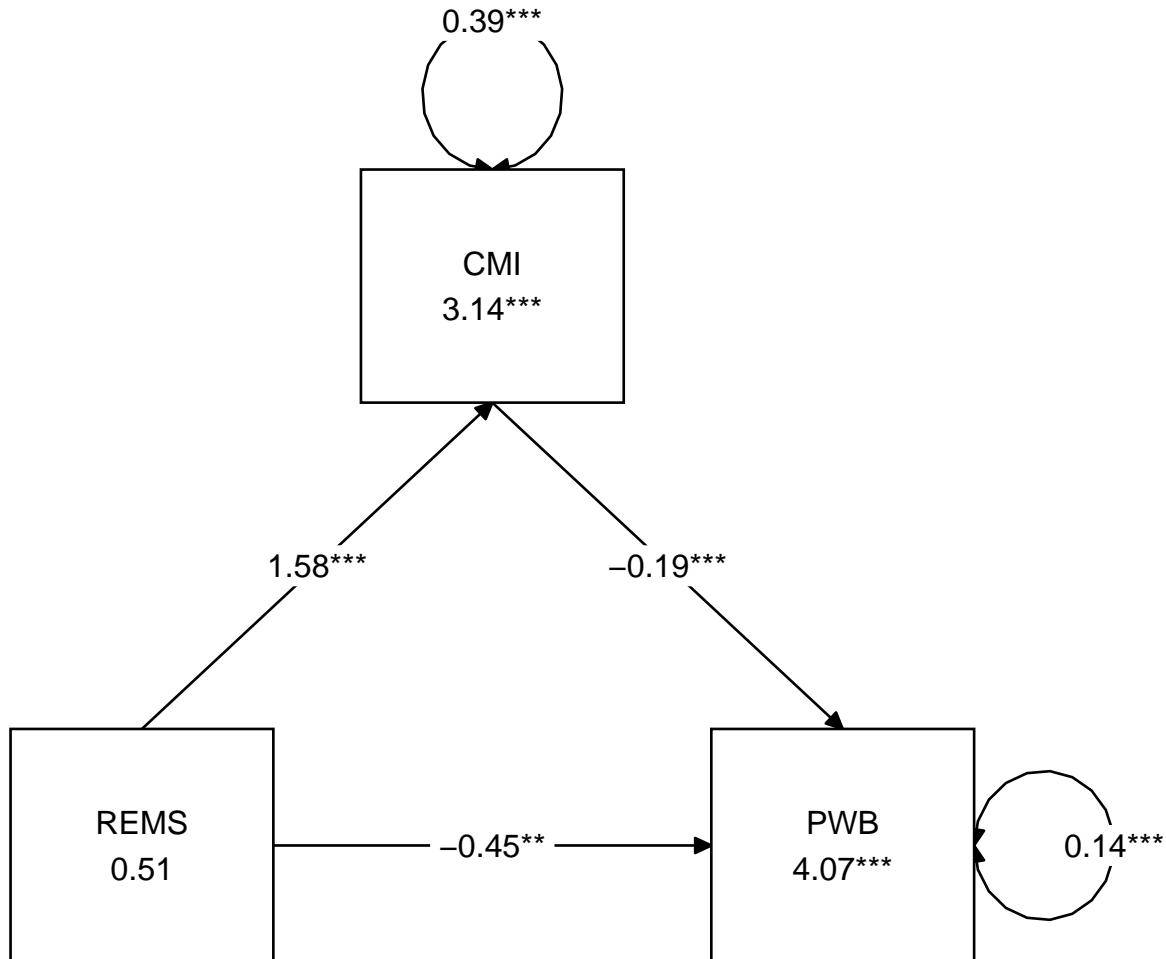
We can write code to remap them

```
med_map2 <- tidySEM::get_layout("", "CMI", "", "REMS", "", "PWB", rows = 2)
med_map2
```

```
##      [,1]  [,2]  [,3]
## [1,] ""    "CMI" ""
## [2,] "REMS" ""    "PWB"
## attr(,"class")
## [1] "layout_matrix" "matrix"      "array"
```

We run again with our map and BOOM! Still needs tinkering for gorgeous, but hey!

```
tidySEM::graph_sem(Kim_fit, layout = med_map2, rect_width = 1.5, rect_height = 1.25,
                     spacing_x = 2, spacing_y = 3, text_size = 4.5)
```



We can use simple code from base R to write the results to a .csv file. This makes it easier to create a table for presenting the results.

```
write.csv(Kim_ParamEsts, file = "KimSimpleMed.csv")
```

Here's how I might organize the data.

Table 2

Model Coefficients Assessing Cultural Mistrust as a Mediator Between Racial Microaggressions and Well-Being

	Cultural Mistrust (M)				Well-Being (Y)			
Antecedent	path	<i>B</i>	<i>SE</i>	<i>p</i>	path	<i>B</i>	<i>SE</i>	<i>p</i>
constant	$i_M$	3.1419	0.103	< 0.001	$i_Y$	4.066	0.184	< 0.001
REMS (X)	$a$	1.576	0.184	< 0.001	$c'$	-0.453	0.136	0.001
CMI (M)					$b$	-0.189	0.052	< 0.001
	$R^2 = 34\%$				$R^2 = 29\%$			

*Note.* The value of the indirect effect was  
 $B = -0.298, SE = 0.093, p = 0.001, 95CI(-0.495, -0.136)$ .

### 5.5.6 Results

A simple mediation model examined the degree to which cultural mistrust mediated the relation of racial microaggressions on well-being. Using the *lavaan* package (v 0.6-16) in R, coefficients for each path, the indirect effect, and total effects were calculated. These values are presented in Table 2 and illustrated in Figure 2. Results suggested that racial/ethnic microaggressions had statistically significant effects on both cultural mistrust ( $B = 1.576, p < 0.001$ ) and well-being ( $B = -0.453, p = 0.001$ ). Further, the indirect effect from our simulated data was statistically significant ( $B = -0.298, SE = 0.093, p = 0.001, 95CI[-0.495, -0.136]$ ). Results suggested that 34% of the variance in cultural mistrust and 29% of the variance in well-being were accounted for by the model.

## 5.6 Considering Covariates

Hayes Chapter 4 [2018] considers the role of covariates (e.g., other variables that could account for some of the variance in the model). When previous research (or commonsense, or detractors) suggest you should include them it is advisable to do so. If they are non-significant and/or your variables continue to explain variance over-and-above their contribution, then you have gained ground in ruling out plausible rival hypotheses and are adding to causal evidence.

Covariates are relatively easy to specify in *lavaan*. I tend to look at my figure and “see where the arrows go.” Those translate readily to the equations we write in the *lavaan* code.

Let’s say we are concerned that anxiety covaries with cultural mistrust and well-being. We’ll add it as a covariate to both.

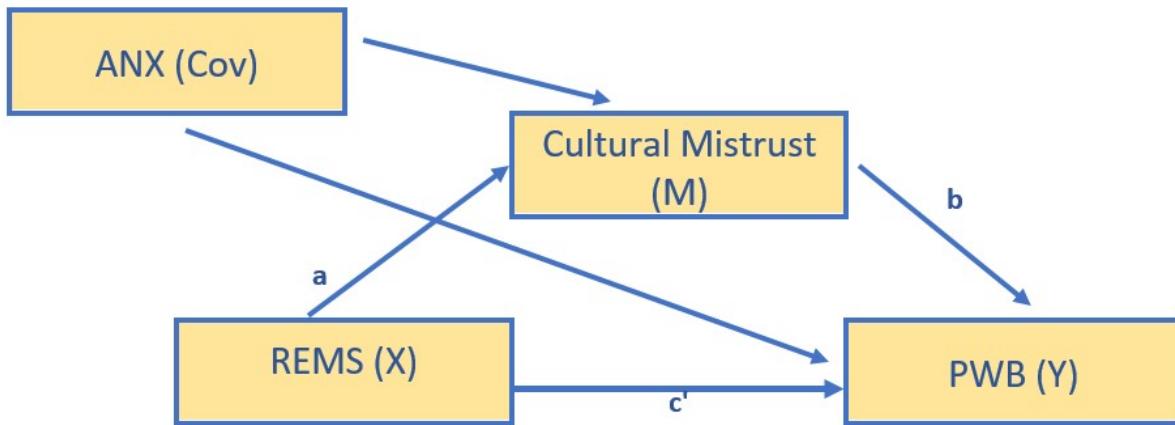


Figure 5.8: Image of the simple mediation model from Kim et al.

```

Kim_fit_covs <- "
  PWB ~ b*CMI + c_p*REMS
  CMI ~a*REMS
  CMI ~ covM*ANX
  PWB ~ covY*ANX

  indirect := a*b
  direct   := c_p
  total_c  := c_p + (a*b)
  "

Kim_fit_covs <- lavaan::sem(Kim_fit_covs, data = dfKim, se = "bootstrap",
  missing = "fiml")
Kcov_sum <- lavaan::summary(Kim_fit_covs, standardized = T, rsq = T, fit = TRUE,
  ci = TRUE)
Kcov_ParEsts <- lavaan::parameterEstimates(Kim_fit_covs, boot.ci.type = "bca.simple",
  standardized = TRUE)
Kcov_sum

## lavaan 0.6.16 ended normally after 1 iteration
##
##      Estimator                               ML
##      Optimization method                     NLMINB
##      Number of model parameters             9
##      Number of observations                 156
##      Number of missing patterns              1
##      Model Test User Model:
##      Test statistic                         0.000
  
```

```

## Degrees of freedom 0
##
## Model Test Baseline Model:
##
## Test statistic 136.009
## Degrees of freedom 5
## P-value 0.000
##
## User Model versus Baseline Model:
##
## Comparative Fit Index (CFI) 1.000
## Tucker-Lewis Index (TLI) 1.000
##
## Robust Comparative Fit Index (CFI) 1.000
## Robust Tucker-Lewis Index (TLI) 1.000
##
## Loglikelihood and Information Criteria:
##
## Loglikelihood user model (H0) -210.170
## Loglikelihood unrestricted model (H1) -210.170
##
## Akaike (AIC) 438.341
## Bayesian (BIC) 465.789
## Sample-size adjusted Bayesian (SABIC) 437.301
##
## Root Mean Square Error of Approximation:
##
## RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: RMSEA <= 0.050 NA
## P-value H_0: RMSEA >= 0.080 NA
##
## Robust RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: Robust RMSEA <= 0.050 NA
## P-value H_0: Robust RMSEA >= 0.080 NA
##
## Standardized Root Mean Square Residual:
##
## SRMR 0.000
##
## Parameter Estimates:
##
## Standard errors Bootstrap
## Number of requested bootstrap draws 1000
## Number of successful bootstrap draws 1000

```

```

## 
## Regressions:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   PWB ~
##     CMI      (b) -0.163  0.053 -3.112  0.002 -0.269 -0.061
##     REMS    (c_p) -0.219  0.150 -1.461  0.144 -0.534  0.062
##   CMI ~
##     REMS      (a)  1.349  0.194  6.948  0.000  0.969  1.711
##     ANX     (covM)  0.198  0.104  1.893  0.058 -0.016  0.398
##   PWB ~
##     ANX     (covY) -0.238  0.063 -3.783  0.000 -0.362 -0.113
##   Std.lv  Std.all
## 
## -0.163 -0.279
## -0.219 -0.139
## 
## 1.349  0.500
## 0.198  0.145
## 
## -0.238 -0.299
## 
## Intercepts:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   .PWB                  4.521  0.213 21.196  0.000  4.088  4.946
##   .CMI                  2.697  0.263 10.267  0.000  2.189  3.268
##   Std.lv  Std.all
##   4.521  10.011
##   2.697  3.497
## 
## Variances:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   .PWB                  0.132  0.015  8.799  0.000  0.101  0.160
##   .CMI                  0.384  0.039  9.760  0.000  0.304  0.457
##   Std.lv  Std.all
##   0.132  0.648
##   0.384  0.645
## 
## R-Square:
##                               Estimate
##   PWB                  0.352
##   CMI                  0.355
## 
## Defined Parameters:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   indirect             -0.220  0.080 -2.769  0.006 -0.381 -0.074
##   direct               -0.219  0.150 -1.460  0.144 -0.534  0.062
##   total_c              -0.440  0.129 -3.419  0.001 -0.702 -0.204
##   Std.lv  Std.all

```

```
##   -0.220  -0.139
##   -0.219  -0.139
##   -0.440  -0.278
```

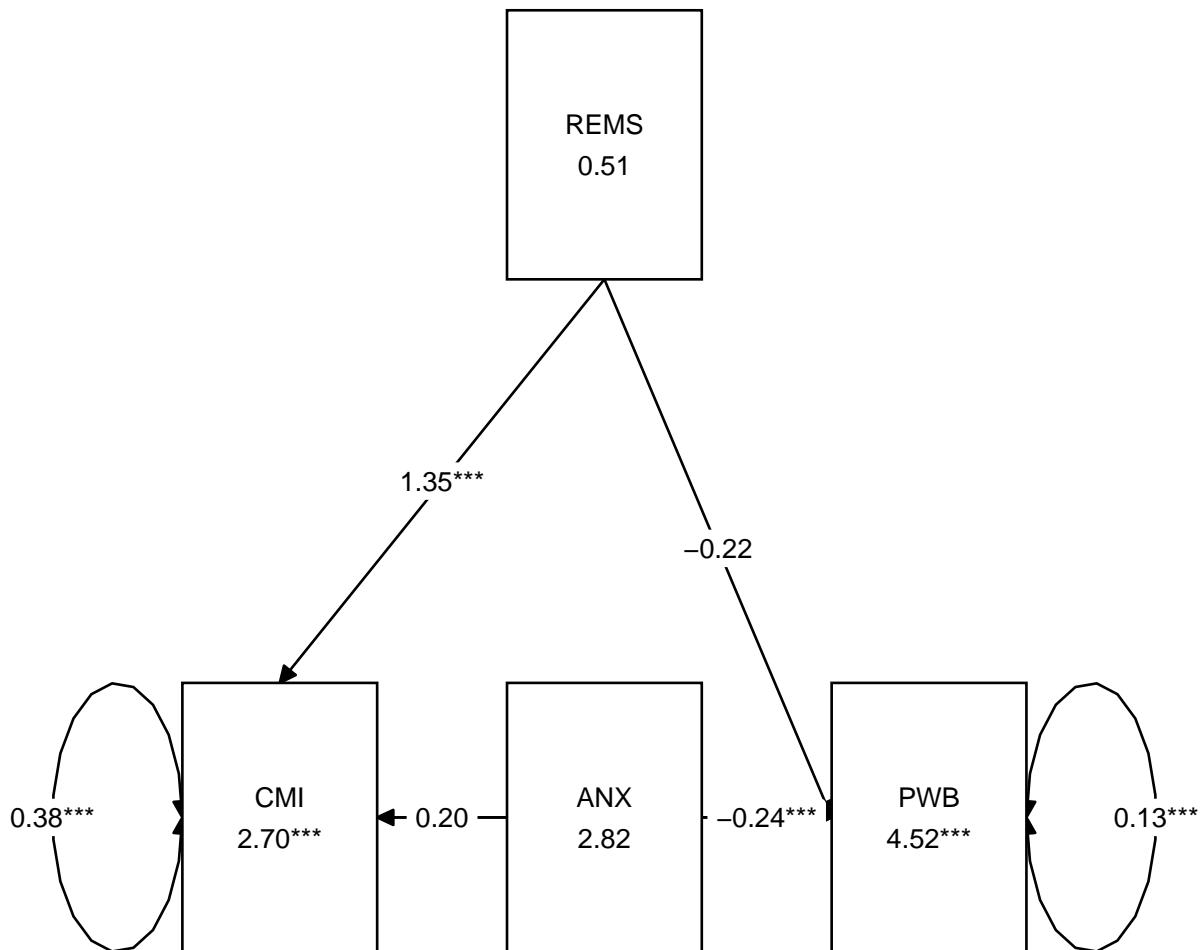
## Kcov\_ParEsts

	lhs	op	rhs	label	est	se	z	pvalue	ci.lower	ci.upper
## 1	PWB	~	CMI	b	-0.163	0.053	-3.112	0.002	-0.261	-0.055
## 2	PWB	~	REMS	c_p	-0.219	0.150	-1.461	0.144	-0.504	0.082
## 3	CMI	~	REMS	a	1.349	0.194	6.948	0.000	0.911	1.692
## 4	CMI	~	ANX	covM	0.198	0.104	1.893	0.058	-0.002	0.407
## 5	PWB	~	ANX	covY	-0.238	0.063	-3.783	0.000	-0.368	-0.114
## 6	PWB	~~	PWB		0.132	0.015	8.799	0.000	0.108	0.170
## 7	CMI	~~	CMI		0.384	0.039	9.760	0.000	0.321	0.478
## 8	REMS	~~	REMS		0.082	0.000	NA	NA	0.082	0.082
## 9	REMS	~~	ANX		0.094	0.000	NA	NA	0.094	0.094
## 10	ANX	~~	ANX		0.320	0.000	NA	NA	0.320	0.320
## 11	PWB	~1			4.521	0.213	21.196	0.000	4.086	4.943
## 12	CMI	~1			2.697	0.263	10.267	0.000	2.180	3.237
## 13	REMS	~1			0.507	0.000	NA	NA	0.507	0.507
## 14	ANX	~1			2.824	0.000	NA	NA	2.824	2.824
## 15	indirect :=		a*b	indirect	-0.220	0.080	-2.769	0.006	-0.386	-0.074
## 16	direct :=		c_p	direct	-0.219	0.150	-1.460	0.144	-0.504	0.082
## 17	total_c :=		c_p+(a*b)	total_c	-0.440	0.129	-3.419	0.001	-0.680	-0.174
	std.lv	std.all	std.nox							
## 1	-0.163	-0.279	-0.279							
## 2	-0.219	-0.139	-0.485							
## 3	1.349	0.500	1.749							
## 4	0.198	0.145	0.256							
## 5	-0.238	-0.299	-0.528							
## 6	0.132	0.648	0.648							
## 7	0.384	0.645	0.645							
## 8	0.082	1.000	0.082							
## 9	0.094	0.580	0.094							
## 10	0.320	1.000	0.320							
## 11	4.521	10.011	10.011							
## 12	2.697	3.497	3.497							
## 13	0.507	1.775	0.507							
## 14	2.824	4.995	2.824							
## 15	-0.220	-0.139	-0.488							
## 16	-0.219	-0.139	-0.485							
## 17	-0.440	-0.278	-0.974							

## 5.6.1 A Figure and a Table

Let's look at a figure to see if we did what we think we did. And to also get a graphic representation of our results.

```
# only worked when I used the library to turn on all these pkgs
library(lavaan)
library(dplyr)
library(ggplot2)
library(tidySEM)
tidySEM::graph_sem(model = Kim_fit_covs)
```



```
tidySEM::get_layout(Kim_fit_covs)
```

```
##      [,1]  [,2]  [,3]
## [1,] NA   "REMS" NA
## [2,] "CMI" "ANX" "PWB"
## attr(,"class")
## [1] "layout_matrix" "matrix"       "array"
```

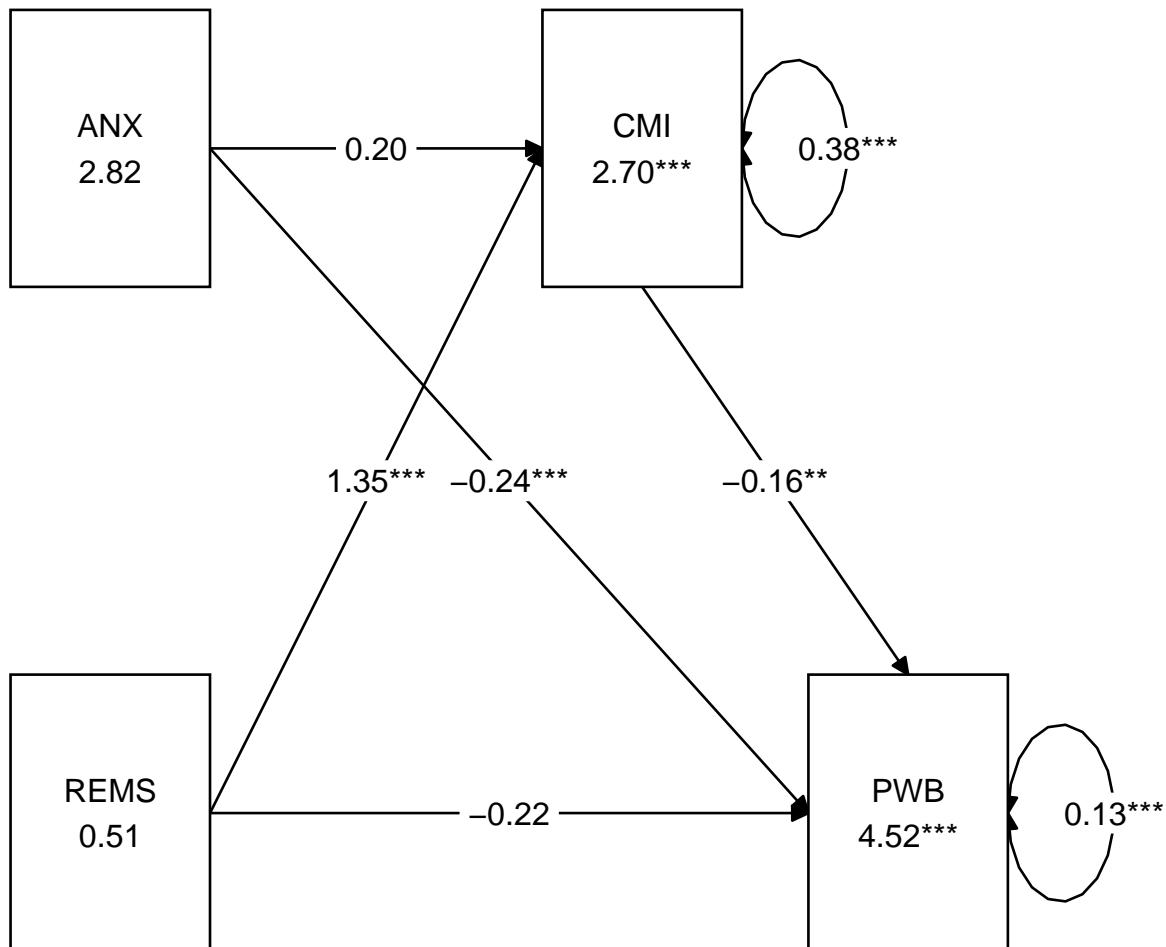
We can write code to remap them

```
med_map3 <- tidySEM::get_layout(
  "ANX", "", "CMI", "",
  "REMS", "", "", "PWB", rows=2)
med_map3

##      [,1]  [,2]  [,3]  [,4]
## [1,] "ANX"  ""    "CMI"  ""
## [2,] "REMS"  ""    ""     "PWB"
## attr(,"class")
## [1] "layout_matrix" "matrix"       "array"
```

We run again with our map and BOOM! Still needs tinkering for gorgeous, but hey!

```
tidySEM::graph_sem(Kim_fit_covs, layout = med_map3, rect_width = 1.5, rect_height = 1.25,
  spacing_x = 2, spacing_y = 3, text_size = 4.5)
```



Below is code to create an outfile that could help with creating a table in a word document or spreadsheet. There will be output that is produced with SEM models that won't be relevant for this project.

```
write.csv(Kcov_ParEsts, file = "KimMedCov.csv")
```

Table 3

---

Model Coefficients Assessing Cultural Mistrust as a Mediator Between Racial Microaggressions and Well-Being

---



---

		Cultural Mistrust (M)				Well-Being (Y)		
Antecedent	path	<i>B</i>	<i>SE</i>	<i>p</i>	path	<i>B</i>	<i>SE</i>	<i>p</i>
constant	$i_M$	2.697	0.256	<0.001	$i_Y$	4.521	0.201	<0.001
REMS (X)	$a$	1.349	0.193	<0.001	$c'$	-0.219	0.143	0.126
CMI (M)					$b$	-0.163	0.050	0.001
ANX (Cov)		0.198	0.193	<0.001		-0.238	0.063	<0.001
$R^2 = 36\%$					$R^2 = 35\%$			

---

*Note.* The value of the indirect effect was  
 $B = -0.220, SE = 0.075, p = 0.004, 95CI(-0.383, -0.090)$ .

---

## 5.6.2 APA Style Write-up

There are varying models for reporting the results of mediation. The Kim et al. [Kim et al., 2017] writeup is a great example. Rather than copying it directly, I have modeled my table after the ones in Hayes [2018] text. You'll notice that information in the table and text are minimally overlapping. APA style cautions us against redundancy in text and table.

### Results

A simple mediation model examined the degree to which cultural mistrust mediated the effect of racial microaggressions on psychological well-being. Using the *lavaan* package (v 0.6-16) in R, coefficients for the each path, the indirect effect, and total effects were calculated. Additionally, the effect of covariate, anxiety, was mapped onto both the mediator and dependent variable. The model accounted for 36% of the variance in cultural mistrust and 35% of the variance in well-being. Supporting the notion of a mediated model, there was a statistically significant indirect effect ( $B = -0.220, SE = 0.075, p = 0.004, 95CI[-0.383, -0.090]$ ) in combination with a non-significant direct effect ( $B = -0.219, p = 0.126$ ) and a statistically significant ( $B = -0.440, p < 0.001$ ).

## 5.7 STAY TUNED

A section on power analysis is planned and coming soon! My apologies that it's not quite Ready.

## 5.8 Residual and Related Questions...

..that you might have; or at least I had, but if had answered them earlier it would have disrupt the flow.

1. Are you sure you can claim a significant indirect effect in the presence of a non-significant total effect? Hayes [2018] is.
  - In the section subtitled, “What about Baron & Kenny” (chapter 4), Hayes argues from both logical/philosophical and statistical perspectives that the size of the total effect does not constrain or determine the size of the indirect effect. That is, an indirect effect can be different from zero even when the total effect is not (pp. 117-119).
2. The output we get is different from the output in the journal article being used as the research vignette. Why? And should we worry about it?
  - We are simulating data. This gives us some advantages in that (unless we specify it), we never have missingness and our variables should be normally distributed. Because we are working from means, standard deviations, and correlations, our data will never be the same as the original researcher. That said, we can compare our results to the journal to *check out work*. In fact, in this very chapter, I got turned around (e.g., first accidentally swapping the mediator and IV; then using the wrong DV) and was able to compare my work against the journal article to correct my errors.
3. Some of the statistics you are reporting are different than the ones in Hayes and the ones that use the PROCESS macro (e.g., what happened to the *F* test)?
  - The default estimator for *lavaan* is maximum likelihood (ML) and Hayes uses ordinary least squares (OLS). This affects both the values of coefficients, standard errors, AND the type of statistics that are reported.
  - You can ask for OLS regression by adding the statement “estimator =”GLS”. Even with this option, I have not discovered a way to obtain the *F* tests for the overall model. Researchers seem to be comfortable with this, even asking for less than we did (e.g., many do not request R square).
  - Best I can tell, researchers who do want this might use a combination of packages, using GLS estimators in *lavaan* (this easily gets them the bootstrapped CIs) and the move to a different regression package to get the intercepts and *F* tests. If I did this I would triple check to make sure that all the output really lined up.
4. Why did we ignore the traditional fit statistics associated with structural equation modeling (e.g., CFI, RMSEA).
  - I hesitate to do this with models that do not include latent variables. Therefore, we asked for an “in-between” amount of info that should be sufficient for publication submission (any editor may have their own preferences and ask for more).

### 5. What if I have missing data?

- When we enter the *lavaan* world we do get options other than multiple imputation. In today's example we used the "sem" fitting function. Unless otherwise specified, listwise deletion (deleting the entire case when one of its variables is used to estimate the model) is the default in *lavaan*. If data are MCAR or MAR, you can add the argument *missing* = "ml" (or its alias *missing* = "fiml"). More here <https://users.ugent.be/~yrosseel/lavaan/lavaan2.pdf> on the 1.7/Missing data in lavaan slide.
- That said, the type of estimator matters. If you estimate your data with GLS (generalized least squares) or WLS (weighted least squares), you are required to have complete data (however you got it). We used maximum likelihood and, even though we had non-missing data, I used the *missing* = "fiml" code.

## 5.9 Practice Problems

The three problems described below are designed to grow with the subsequent chapters on complex mediation and conditional process analysis (i.e., moderated mediation). Therefore, I recommend that you select a dataset that includes at least four variables. If you are new to this topic, you may wish to select variables that are all continuously scaled. The IV and moderator (subsequent chapters) could be categorical (if they are dichotomous, please use 0/1 coding; if they have more than one category it is best if they are ordered). You will likely encounter challenges that were not covered in this chapter. Search for and try out solutions, knowing that there are multiple paths through the analysis.

The suggested practice problem for this chapter is to conduct a simple mediation.

### 5.9.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed

If this topic feels a bit overwhelming, simply change the random seed in the data simulation, then rework the problem. This should provide minor changes to the data (maybe in the second or third decimal point), but the results will likely be very similar.

### 5.9.2 Problem #2: Rework the research vignette, but swap one or more variables

Use the simulated data, but select one of the other models that was evaluated in the Kim et al. [2017] study. Compare your results to those reported in the manuscript.

### 5.9.3 Problem #3: Use other data that is available to you

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete a simple mediation.

### 5.9.4 Grading Rubric

Assignment Component	Points Possible	Points Earned
1. Assign each variable to the X, Y, or M roles (ok but not required to include a cov)	5	_____
2. Import the data and format the variables in the model	5	_____
3. Specify and run the lavaan model	5	_____
4. Use tidySEM to create a figure that represents your results	5	_____
5. Create a table that includes regression output for the M and Y variables	5	_____
6. Represent your work in an APA-style write-up	5	_____
7. Explanation to grader	5	_____
8. Be able to hand-calculate the indirect, direct, and total effects from the a, b, & c' paths	5	_____
<b>Totals</b>	<b>35</b>	_____

## 5.10 Homeworked Example

### Screencast Link

For more information about the data used in this homeworked example, please refer to the description and codebook located at the end of the [introductory lesson](#) in [ReCentering Psych Stats](#). An .rds file which holds the data is located in the [Worked Examples](#) folder at the GitHub site the hosts the OER. The file name is *ReC.rds*.

The suggested practice problem for this chapter is to conduct a simple mediation.

### 5.10.1 Assign each variable to the X, Y, or M roles (ok but not required to include a covariate)

X = Centering: explicit recentering (0 = precentered; 1 = recentered) M = TradPed: traditional pedagogy (continuously scaled with higher scores being more favorable) Y = SRPed: socially responsive pedagogy (continuously scaled with higher scores being more favorable)

### Specify a research model

I am hypothesizing that the evaluation of social responsive pedagogy is predicted by intentional recentering through traditional pedagogy.

### Import the data and format the variables in the model

```
raw <- readRDS("ReC.rds")
```

I need to score the TradPed and SRPed variables

```
TradPed_vars <- c("ClearResponsibilities", "EffectiveAnswers", "Feedback",
  "ClearOrganization", "ClearPresentation")
raw$TradPed <- sjstats::mean_n(raw[, ..TradPed_vars], 0.75)

SRPed_vars <- c("InclusvClassrm", "EquitableEval", "MultPerspectives",
  "DEIintegration")
raw$SRPed <- sjstats::mean_n(raw[, ..SRPed_vars], 0.75)
```

I will create a babydf.

```
babydf <- dplyr::select(raw, Centering, TradPed, SRPed)
```

Let's check the structure of the variables:

```
str(babydf)
```

### Specify and run the lavaan model

```
ReCMed <- "
  SRPed ~ b*TradPed + c_p*Centering
  TradPed ~ a*Centering

  indirect := a*b
  direct   := c_p
  total_c  := c_p + (a*b)
  "

ReCfit <- lavaan::sem(ReCMed, data = babydf, se = "bootstrap", missing = "fiml")
ReCsummary <- lavaan::summary(ReCfit, standardized = T, rsq = T, fit = TRUE,
  ci = TRUE)
ReC_ParamEsts <- lavaan::parameterEstimates(ReCfit, boot.ci.type = "bca.simple",
  standardized = TRUE)
ReCsummary
```

```
## lavaan 0.6.16 ended normally after 14 iterations
##
##   Estimator                               ML
##   Optimization method                    NLMINB
##   Number of model parameters             7
##   Number of observations                  310
##   Number of missing patterns              4
##
```

```

## Model Test User Model:
##
## Test statistic 0.000
## Degrees of freedom 0
##
## Model Test Baseline Model:
##
## Test statistic 216.492
## Degrees of freedom 3
## P-value 0.000
##
## User Model versus Baseline Model:
##
## Comparative Fit Index (CFI) 1.000
## Tucker-Lewis Index (TLI) 1.000
##
## Robust Comparative Fit Index (CFI) 1.000
## Robust Tucker-Lewis Index (TLI) 1.000
##
## Loglikelihood and Information Criteria:
##
## Loglikelihood user model (H0) -506.434
## Loglikelihood unrestricted model (H1) -506.434
##
## Akaike (AIC) 1026.868
## Bayesian (BIC) 1053.024
## Sample-size adjusted Bayesian (SABIC) 1030.823
##
## Root Mean Square Error of Approximation:
##
## RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: RMSEA <= 0.050 NA
## P-value H_0: RMSEA >= 0.080 NA
##
## Robust RMSEA 0.000
## 90 Percent confidence interval - lower 0.000
## 90 Percent confidence interval - upper 0.000
## P-value H_0: Robust RMSEA <= 0.050 NA
## P-value H_0: Robust RMSEA >= 0.080 NA
##
## Standardized Root Mean Square Residual:
##
## SRMR 0.000
##
## Parameter Estimates:
##

```

```

## Standard errors                                Bootstrap
## Number of requested bootstrap draws          1000
## Number of successful bootstrap draws         1000
##
## Regressions:
##                         Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## SRPed ~
##   TradPed    (b)      0.549    0.046  11.891  0.000   0.454   0.639
##   Centerng (c_p)    0.127    0.047   2.690  0.007   0.041   0.229
## TradPed ~
##   Centerng (a)     -0.101    0.085  -1.193  0.233  -0.272   0.066
## Std.lv Std.all
##
##   0.549    0.716
##   0.127    0.107
##
##   -0.101   -0.066
##
## Intercepts:
##                         Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .SRPed           2.006    0.234   8.591  0.000   1.523   2.459
## .TradPed         4.394    0.129  34.082  0.000   4.129   4.648
## Std.lv Std.all
##   2.006    3.440
##   4.394    5.778
##
## Variances:
##                         Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .SRPed           0.165    0.018   9.053  0.000   0.130   0.199
## .TradPed         0.576    0.074   7.789  0.000   0.434   0.724
## Std.lv Std.all
##   0.165    0.486
##   0.576    0.996
##
## R-Square:
##                         Estimate
## SRPed            0.514
## TradPed          0.004
##
## Defined Parameters:
##                         Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## indirect        -0.056    0.048  -1.151  0.250  -0.159   0.033
## direct          0.127    0.047   2.689  0.007   0.041   0.229
## total_c         0.071    0.066   1.072  0.284  -0.055   0.205
## Std.lv Std.all
##   -0.056   -0.047
##   0.127    0.107
##   0.071    0.060

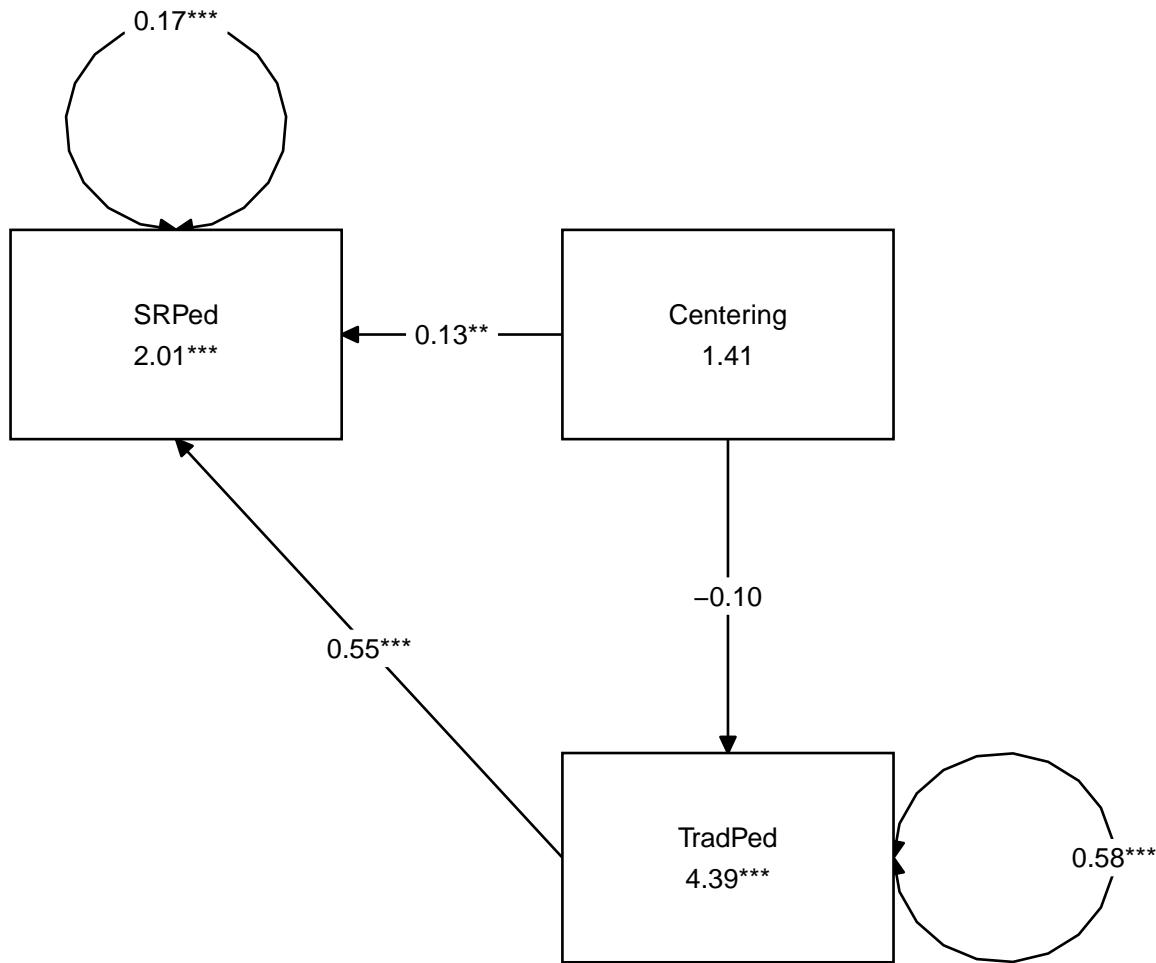
```

```
ReC_ParamEsts
```

##	lhs	op	rhs	label	est	se	z	pvalue	ci.lower	ci.upper
## 1	SRPed	~	TradPed	b	0.549	0.046	11.891	0.000	0.454	0.639
## 2	SRPed	~	Centering	c_p	0.127	0.047	2.690	0.007	0.034	0.223
## 3	TradPed	~	Centering	a	-0.101	0.085	-1.193	0.233	-0.272	0.066
## 4	SRPed	~~	SRPed		0.165	0.018	9.053	0.000	0.133	0.202
## 5	TradPed	~~	TradPed		0.576	0.074	7.789	0.000	0.440	0.732
## 6	Centering	~~	Centering		0.241	0.000	NA	NA	0.241	0.241
## 7	SRPed	~1			2.006	0.234	8.591	0.000	1.528	2.460
## 8	TradPed	~1			4.394	0.129	34.082	0.000	4.127	4.640
## 9	Centering	~1			1.406	0.000	NA	NA	1.406	1.406
## 10	indirect	:=	a*b	indirect	-0.056	0.048	-1.151	0.250	-0.159	0.033
## 11	direct	:=	c_p	direct	0.127	0.047	2.689	0.007	0.034	0.223
## 12	total_c	:=	c_p+(a*b)	total_c	0.071	0.066	1.072	0.284	-0.060	0.200
##	std.lv	std.all	std.nox							
## 1	0.549	0.716	0.716							
## 2	0.127	0.107	0.217							
## 3	-0.101	-0.066	-0.133							
## 4	0.165	0.486	0.486							
## 5	0.576	0.996	0.996							
## 6	0.241	1.000	0.241							
## 7	2.006	3.440	3.440							
## 8	4.394	5.778	5.778							
## 9	1.406	2.863	1.406							
## 10	-0.056	-0.047	-0.096							
## 11	0.127	0.107	0.217							
## 12	0.071	0.060	0.122							

Use tidySEM to create a figure that represents your results

```
# only worked when I used the library to turn on all these pkgs
library(lavaan)
library(dplyr)
library(ggplot2)
library(tidySEM)
tidySEM::graph_sem(model = ReCfit)
```

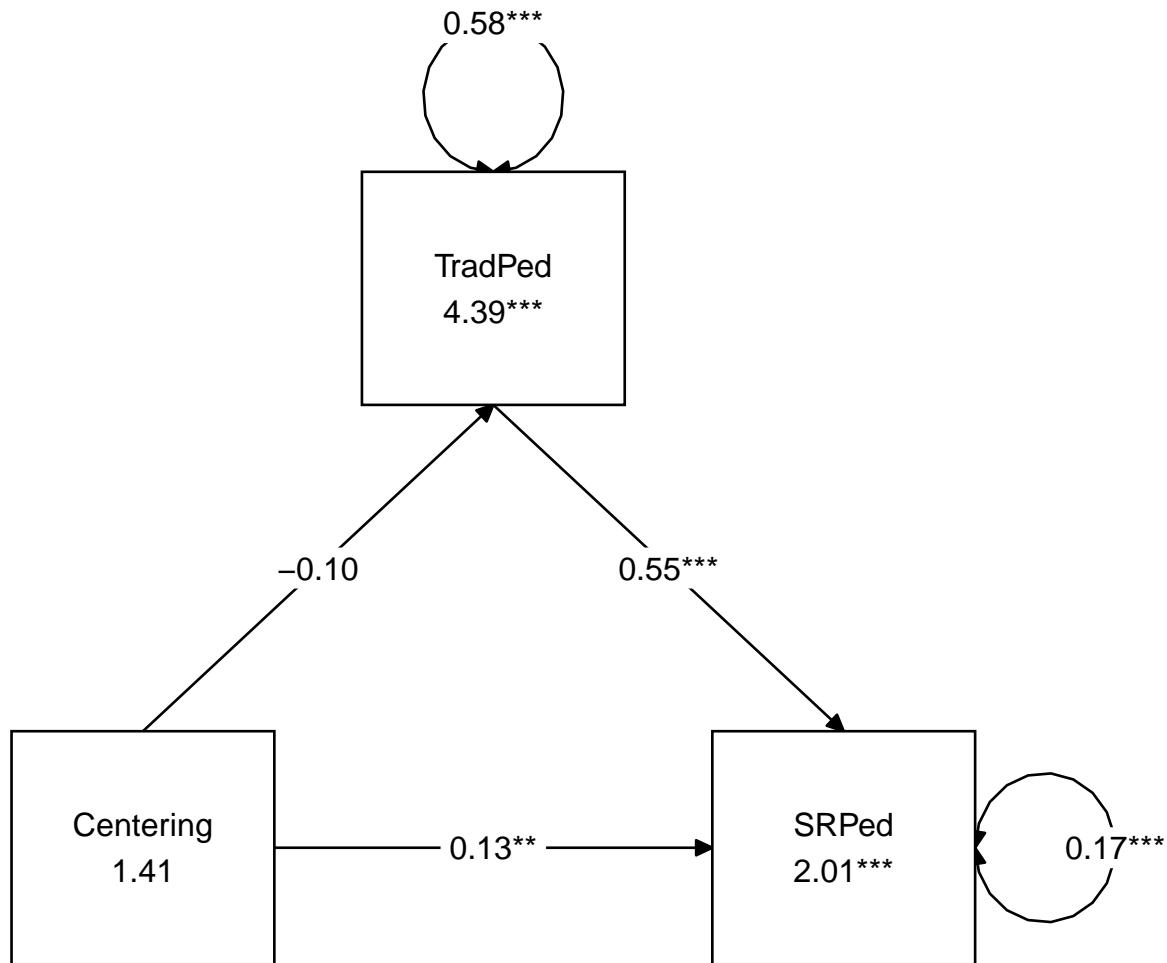


`tidy=TRUE, tidy.opts=list(width.cutoff=70)}` tidySEM::get\_layout(ReCfit) We can write code to remap them

```
med_map <- tidySEM::get_layout("", "TradPed", "", "Centering", "", "SRPed",
  rows = 2)
med_map
```

```
##      [,1]      [,2]      [,3]
## [1,] ""      "TradPed"  ""
## [2,] "Centering"  ""      "SRPed"
## attr(),"class")
## [1] "layout_matrix" "matrix"      "array"
```

```
tidySEM::graph_sem(ReCfit, layout=med_map, rect_width = 1.5, rect_height = 1.25, spacing_x = 2)
```



Create a table that includes regression output for the M and Y variables

```
write.csv(ReC_ParamEsts, file = "ReCSimpMed.csv")
```

Table 1

## Model Coefficients Assessing Traditional Pedagogy as a Mediator Between Centering and Socially Responsive Pedagogy

Traditional Pedagogy (M)					Socially Responsive Pedagogy (Y)				
Antecedent	path	B	SE	p	path	B	SE	p	
Antecedent	path	B	SE	p	path	B	SE	p	

constant	$i_M$	4.394	0.133	< 0.001	$i_Y$	2.006	0.238	< 0.001
Centering	$a$	-0.101	0.088	< 0.247	$c'$	0.127	0.048	0.008
(X)								
TradPed					$b$	0.549	0.047	< 0.001
(M)								
				$R^2 = 0.4\%$				
					$R^2 = 51\%$			

*Note.* Centering: 0 = pre-centered, 1 = recentered. TradPed is traditional pedagogy. The value of the indirect effect was  $B = -0.056, SE = 0.051, p = 0.272, 95CI(-0.163, 0.035)$

### Represent your work in an APA-style write-up

A simple mediation model examined the degree to which evaluations of traditional pedagogy mediated the relation of explicit recentering on socially responsive pedagogy. Using the *lavaan* package (v 0.6-16) in R, coefficients for each path, the indirect effect, and total effects were calculated. These values are presented in Table 1 and illustrated in Figure 1. Results suggested that negligible (.4%) of the variance was accounted for in traditional pedagogy. In contrast 51% of the variance was accounted for in socially responsive pedagogy. The indirect effect ( $B = -0.056, SE = 0.051, p = 0.272, 95CI[-0.163, 0.035]$ ) was statistically significant. Comparing total and direct effects, the total effect of centering and traditional pedagogy on socially responsive pedagogy was not statistically significant ( $B = 0.071, p = 0.302$ ). In contrast, the direct effect was ( $B = 0.127, p = 0.008$  was not). This suggests that while centering and traditional pedagogy do influence socially responsive pedagogy, their influence is relatively independent.

```
apaTables::apa.cor.table(babydf, table.number = 1, show.sig.stars = TRUE,
  landscape = TRUE, filename = NA)

##
## Table 1
##
## Means, standard deviations, and correlations with confidence intervals
##
##
##      Variable   M     SD    1
## 1. TradPed 4.25  0.76
##
## 2. SRPed   4.52  0.58 .71**
##                   [.65, .76]
##
##
## Note. M and SD are used to represent mean and standard deviation, respectively.
```

```
## Values in square brackets indicate the 95% confidence interval.
## The confidence interval is a plausible range of population correlations
## that could have caused the sample correlation (Cumming, 2014).
## * indicates p < .05. ** indicates p < .01.
##
```

### Explanation to grader

Be able to hand-calculate the indirect, direct, and total effects from the a, b, & c' paths

- Indirect =  $a^*b$
- Direct = Total minus indirect
- Total =  $(a^*b) + c'$

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/Los_Angeles
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics   grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] lubridate_1.9.2 forcats_1.0.0 stringr_1.5.0    purrr_1.0.1
## [5] readr_2.1.4     tidyr_1.3.0    tibble_3.2.1    tidyverse_2.0.0
## [9] tidySEM_0.2.4   OpenMx_2.21.8  ggplot2_3.4.3   dplyr_1.1.2
## [13] lavaan_0.6-16  psych_2.3.6
##
## loaded via a namespace (and not attached):
## [1] mnormt_2.1.1        gridExtra_2.3       formatR_1.14
## [4] inline_0.3.19       sandwich_3.0-2      rlang_1.1.1
## [7] magrittr_2.0.3       multcomp_1.4-25     matrixStats_1.0.0
```

```

## [10] compiler_4.3.1          loo_2.6.0           callr_3.7.3
## [13] vctrs_0.6.3             quadprog_1.5-8    pkgconfig_2.0.3
## [16] crayon_1.5.2            fastmap_1.1.1     backports_1.4.1
## [19] bain_0.2.8              labeling_0.4.2    pbivnorm_0.6.0
## [22] pander_0.6.5            utf8_1.2.3        rmarkdown_2.24
## [25] tzdb_0.4.0              nloptr_2.0.3      ps_1.7.5
## [28] xfun_0.39                highr_0.10       sjmisc_2.8.9
## [31] broom_1.0.5              parallel_4.3.1   prettyunits_1.1.1
## [34] R6_2.5.1                 stringi_1.7.12   StanHeaders_2.26.27
## [37] parallelly_1.36.0        car_3.1-2        boot_1.3-28.1
## [40] estimability_1.4.1       Rcpp_1.0.10      bookdown_0.34
## [43] rstan_2.21.8             knitr_1.43       modelr_0.1.11
## [46] future.apply_1.11.0       zoo_1.8-12       bayesplot_1.10.0
## [49] splines_4.3.1            timechange_0.2.0 Matrix_1.5-4.1
## [52] igraph_1.5.1              tidyselect_1.2.0 rstudioapi_0.15.0
## [55] abind_1.4-5              yaml_2.3.7       sjlabelled_1.2.0
## [58] codetools_0.2-19          tmvnsim_1.0-2    processx_3.8.1
## [61] listenv_0.9.0             pkgbuild_1.4.2   lattice_0.21-8
## [64] nonnest2_0.5-5            plyr_1.8.8       bayestestR_0.13.1
## [67] withr_2.5.0              coda_0.19-4      evaluate_0.21
## [70] survival_3.5-5           future_1.33.0   fastDummies_1.7.3
## [73] CompQuadForm_1.4.3        RcppParallel_5.1.7 texreg_1.38.6
## [76] pillar_1.9.0              carData_3.0-5    checkmate_2.2.0
## [79] stats4_4.3.1              insight_0.19.3   generics_0.1.3
## [82] dbscan_1.1-11             hms_1.1.3        rstantools_2.3.1
## [85] munsell_0.5.0              scales_1.2.1     blavaan_0.4-8
## [88] minqa_1.2.5              globals_0.16.2   xtable_1.8-4
## [91] glue_1.6.2                 emmeans_1.8.7    tools_4.3.1
## [94] data.table_1.14.8         lme4_1.1-33     gsubfn_0.7
## [97] RANN_2.6.1                 mvtnorm_1.2-2    grid_4.3.1
## [100] MplusAutomation_1.1.0     apaTables_2.0.8   colorspace_2.1-0
## [103] nlme_3.1-162              performance_0.10.4 proto_1.0.0
## [106] cli_3.6.1                 fansi_1.0.4      sjstats_0.18.2
## [109] gtable_0.3.3              digest_0.6.32   progressr_0.13.0
## [112] TH.data_1.1-2             farver_2.1.1     htmltools_0.5.5
## [115] lifecycle_1.0.3            httr_1.4.7       MASS_7.3-60

```



# Chapter 6

## Complex Mediation

### [Screencasted Lecture Link](#)

The focus of this chapter is the extension of simple mediation to models with multiple mediators. In these models with greater complexity we look at both parallel and serial mediation. There is also more elaboration on some of the conceptual issues related to the estimation of indirect effects.

### 6.1 Navigating this Lesson

There is about 1 hour and 20 minutes of lecture. If you work through the materials with me it would be plan for an additional two hours.

While the majority of R objects and data you will need are created within the R script that sources the chapter, there are a few that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

#### 6.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Define *epiphénomérité* and explain how it is related to (and supports the notion of) multiple mediation.
- Distinguish between parallel and serial mediation models.
- Locate and interpret *lavaan* output from multiply mediated models including
  - identifying coefficients,
  - percentage of variance accounted for,
- all the effects (total, direct, indirect, total indirect),
- contrasts (comparing the significance of the indirect effects).
- Explain the limitations of the classic approach [[Baron and Kenny, 1986](#)] to mediation.

### 6.1.2 Planning for Practice

The suggestions for practice in this chapter include conducting parallel, serial, and/or mediation models. Options of graded complexity could include:

- Rework the problem in the chapter by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- There are a number of variables in the dataset that sourced the research vignettes for this and the prior chapter on **simple mediation**. Swap out one or more variables in a parallel or serial (or both) model.
- Conduct a parallel or serial (or both) mediation with data to which you have access. This could include data you simulate on your own or from a published article.

### 6.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press. Available as an ebook from the SPU library: [https://alliance-spu.primo.exlibrisgroup.com/permalink/01ALLIANCE\\_SPU/1q85832/alma99900435260301847](https://alliance-spu.primo.exlibrisgroup.com/permalink/01ALLIANCE_SPU/1q85832/alma99900435260301847)
  - **Chapter 5: More than One Mediator:** This chapter walks the reader through parallel and serial mediation models. We will do both!
  - **Appendix A: Using Process:** An essential tool for PROCESS users because, even when we are in the R environment, this is the “idea book.” That is, the place where all the path models are presented in figures.
- Lewis, J. A., Williams, M. G., Peppers, E. J., & Gadson, C. A. (2017). Applying intersectionality to explore the relations between gendered racism and health among Black women. *Journal of Counseling Psychology*, 64(5), 475–486. <https://doi-org.ezproxy.spu.edu/10.1037/cou0000231>

### 6.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
if (!require(lavaan)) {
  install.packages("lavaan")
}
if (!require(tidyverse)) {
  install.packages("tidyverse")
}
```

```

if (!require(dplyr)) {
  install.packages("dplyr")
}
if (!require(psych)) {
  install.packages("psych")
}
if (!require(apaTables)) {
  install.packages("apaTables")
}
if (!require(tidySEM)) {
  install.packages("tidySEM")
}

```

## 6.2 Complex Mediation

The simple mediation model is quite popular, but also limiting in that it:

- frequently oversimplifies the processes we want to study, and
- is likely mis-specified, in that there are unmodeled mechanisms.

Hayes [2022b] identified four reasons to consider multiply mediated models:

- We are generally interested in MULTIPLE mechanisms
- A mechanism (such as a mediator) in the model, might, itself be mediated (i.e., mediated mediation)
- *Epiphenomenality* (“unknown confounds”): a proposed mediator could be related to an outcome not because it causes the outcome, but because it is correlated with another variable that is causally influencing the outcome. This is a noncausal alternative explanation for an association.
- Including multiple mediators allows formal comparison of the strength of the mediating mechanisms.

There are two multiple mediator models that we will consider: parallel, serial.

## 6.3 Workflow for Complex Mediation

The following is a proposed workflow for conducting a complex mediation.

Conducting a parallel or serial (i.e., complex) mediation involves the following steps:

1. Conducting an a priori power analysis to determine the appropriate sample size.
  - This will require estimates of effect that are drawn from pilot data, the literature, or both.

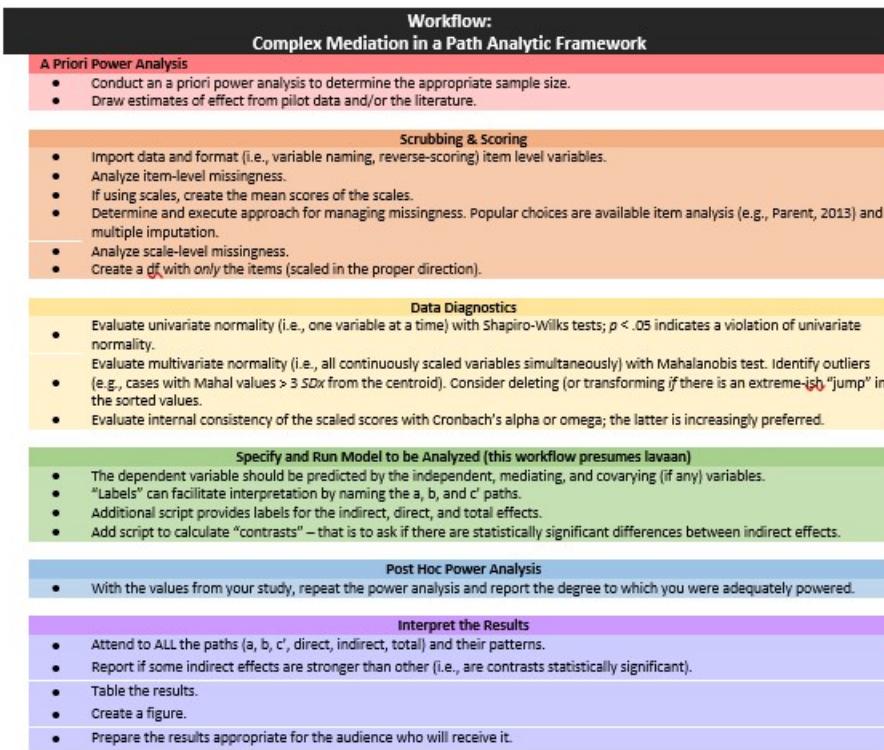


Figure 6.1: A colorful image of a workflow for complex mediation

2. **Scrubbing** and **scoring** the data.
  - Guidelines for such are presented in the respective lessons.
3. Conducting data diagnostics, this includes:
  - item and scale level missingness,
  - internal consistency coefficients (e.g., alphas or omegas) for scale scores,
  - univariate and multivariate normality
4. Specifying and running the model (this lesson presumes it will with the R package, *lavaan*).
  - The dependent variable should be predicted by the independent, mediating, and covarying (if any) variables.
  - "Labels" can facilitate interpretation by naming the a, b, and c' paths.
  - Additional script provides labels for the indirect, direct, and total effects.
  - With multiple indirect effects, specify contrasts to see if they are statistically significantly different form each other.
5. Conducting a post hoc power analysis.
  - Informed by your own results, you can see if you were adequately powered to detect a statistically significant effect, if, in fact, one exists.
6. Interpret and report the results.
  - Interpret ALL the paths and their patterns.

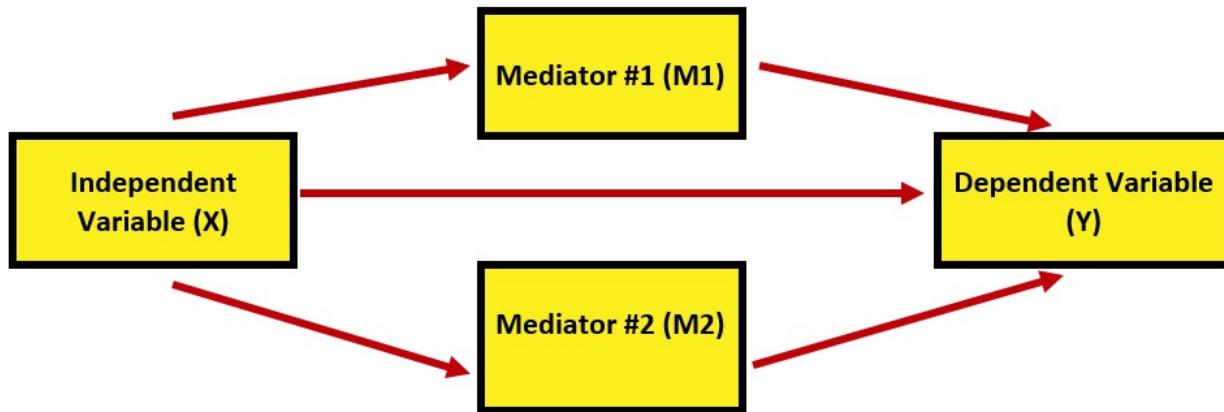
- Report if some indirect effects are stronger than others (i.e., results of the contrasts).
- Create a table and figure.
- Prepare the results in a manner that is useful to your audience.

## 6.4 Parallel Mediation

**Parallel multiple mediation:** An antecedent variable X is modeled as influencing consequent Y directly as well as indirectly through two or more mediators, with the condition that no mediator causally influences another [Hayes, 2022b, p. 161]

With multiple mediation we introduce additional effects:

- *Direct effect,  $c'$*  (this is not new) quantifies how much two cases that differ by a unit on X are estimated to differ on Y – independent of all mediators.
- *Specific indirect effect,  $a_i b_i$* , the individual mediated effects
- *Total indirect effects*,  $\sum_{i=1}^k a_i b_i$  the sum of the values of the specific indirect effects. The total indirect effect can also be calculated by subtracting the direct effects from the total effects:  $c - c'$
- *Total effect of X on Y,  $c = c' + \sum_{i=1}^k a_i b_i$*  (also not new) the sum of the direct and indirect effects. The total effect can also be estimated by regressing Y on X alone.
- *Contrasts* allow us to directly compare separate mediating effects to see if one indirect effect is stronger than the other.



In this parallel model, we can describe these effects this way:

- *Direct effect:* The effect of IV on the DV, accounting for two mediators (indirect effects) in the model.
- *Specific indirect effects:* There are indirect (or mediating) paths from the IV to the DV; through M1 and M2, respectively.
- *Total indirect effect of X on Y:* A sum of the value of indirect effects through the specific indirect effects (M1 and M2).

- *Total effect*: The sum of the direct and indirect effects. Also calculated by regressing Y (dependent variable) on X (independent variable) alone, without any other variables in the model.

Recall that for a complex mediation to be parallel, there can be no causal links between mediators. This is true in this example.

### 6.4.1 A Mechanical Example

Let's work a mechanical example with simulated data that assures a statistically significant outcome. Credit to this example is from the Paulo Toffanin website [[Toffanin, 2017](#)].

We can bake our own data by updating the script we used in simple mediation to add a second mediator.

#### 6.4.1.1 Data Simulation

```
# Concerned that identical variable names across book chapters may be
# problematic, I'm adding 'p' in front the 'Data' variable.
set.seed(230925)
X <- rnorm(100)
M1 <- 0.5 * X + rnorm(100)
M2 <- -0.35 * X + rnorm(100)
Y <- 0.7 * M2 + 0.48 * M1 + rnorm(100)
pData <- data.frame(X = X, Y = Y, M1 = M1, M2 = M2)
```

Using what we learned in conducting a simple mediation in *lavaan*, we can look at the figure of our proposed model and *backwardstrace* the paths to write the code.

Remember...

- The model exists between 2 single quotation marks (the odd looking ' and ' at the beginning and end).
- You can write the Y as I have done in the R chunk below, or you can write the Y separately from each arrow, such as
  - $Y \sim b1*M1$
  - $Y \sim b2*M2$
  - $Y \sim c_p*X$
- Everything else transfers from our simple mediation, remember that
  - the asterisk (“\*”) allows us to assign labels (a1, a2, b1, b2, etc.) to the paths; these are helpful for intuitive interpretation
  - that eyes/nose notation (:=) is used when creating a new variable that is a function of variables in the model, but not in the dataset (i.e., the a and b path).

- in traditional mediation speak, the direct path from X to Y is  $c'$  ( $c$  prime) and the total effect of X to Y (with nothing else in the model) is just  $c$ . Hence the  $c_p$  label for  $c$  prime.
- Something new: the *contrast* statement (only one in this example, but you could have more) allows us to compare the indirect effects to each other. We specify it in the lavaan model, but then need to test it in a subsequent set of script.
- *Note:* In the online example, the writer adds code to correlate M1 and M2. This didn't/doesn't seem right to me and then, later, when we amend it to be a serial model, it made even less sense to have them be correlated.

#### 6.4.1.2 Specifying lavaan code

```

parallel_med <- "
  Y ~ b1*M1 + b2*M2 + c_p*X
  M1 ~ a1*X
  M2 ~ a2*X

  indirect1 := a1 * b1
  indirect2 := a2 * b2
  contrast := indirect1 - indirect2
  total_indirects := indirect1 + indirect2
  total_c      := c_p + (indirect1) + (indirect2)
  direct := c_p
"

parallel_fit <- lavaan::sem(parallel_med, data = pData, se = "bootstrap",
  missing = "fiml", bootstrap = 1000)
pfit_sum <- lavaan::summary(parallel_fit, standardized = TRUE, rsq = T,
  fit = TRUE, ci = TRUE)
pfit_ParEsts <- lavaan::parameterEstimates(parallel_fit, boot.ci.type = "bca.simple",
  standardized = TRUE)
pfit_sum

## lavaan 0.6.16 ended normally after 1 iteration
##
##           Estimator          ML
##           Optimization method    NLINMB
##           Number of model parameters   11
##           Number of observations     100
##           Number of missing patterns     1
##           Model Test User Model:
##           Test statistic        2.475

```

```

## Degrees of freedom                                1
## P-value (Chi-square)                            0.116
##
## Model Test Baseline Model:
##
## Test statistic                                 126.642
## Degrees of freedom                           6
## P-value                                      0.000
##
## User Model versus Baseline Model:
##
## Comparative Fit Index (CFI)                  0.988
## Tucker-Lewis Index (TLI)                      0.927
##
## Robust Comparative Fit Index (CFI)            0.988
## Robust Tucker-Lewis Index (TLI)                0.927
##
## Loglikelihood and Information Criteria:
##
## Loglikelihood user model (H0)                 -433.660
## Loglikelihood unrestricted model (H1)          -432.423
##
## Akaike (AIC)                                    889.321
## Bayesian (BIC)                                  917.977
## Sample-size adjusted Bayesian (SABIC)           883.237
##
## Root Mean Square Error of Approximation:
##
## RMSEA                                         0.121
## 90 Percent confidence interval - lower        0.000
## 90 Percent confidence interval - upper         0.322
## P-value H_0: RMSEA <= 0.050                  0.161
## P-value H_0: RMSEA >= 0.080                  0.772
##
## Robust RMSEA                                    0.121
## 90 Percent confidence interval - lower        0.000
## 90 Percent confidence interval - upper         0.322
## P-value H_0: Robust RMSEA <= 0.050           0.161
## P-value H_0: Robust RMSEA >= 0.080           0.772
##
## Standardized Root Mean Square Residual:
##
## SRMR                                         0.046
##
## Parameter Estimates:
##
## Standard errors                               Bootstrap
## Number of requested bootstrap draws          1000

```

```

## Number of successful bootstrap draws 1000
##
## Regressions:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## Y ~
##   M1      (b1)    0.456   0.111   4.123   0.000   0.247   0.670
##   M2      (b2)    0.743   0.074  10.095   0.000   0.611   0.903
##   X       (c_p)    0.030   0.100   0.301   0.764  -0.176   0.214
## M1 ~
##   X       (a1)    0.510   0.079   6.480   0.000   0.357   0.673
## M2 ~
##   X       (a2)   -0.381   0.121  -3.152   0.002  -0.619  -0.135
## Std.lv Std.all
##
##   0.456   0.383
##   0.743   0.693
##   0.030   0.025
##
##   0.510   0.502
##
##   -0.381  -0.338
##
## Intercepts:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .Y        0.113   0.092   1.224   0.221  -0.068   0.289
## .M1     -0.089   0.097  -0.913   0.361  -0.262   0.113
## .M2      0.017   0.120   0.140   0.888  -0.215   0.256
## Std.lv Std.all
##   0.113   0.083
##   -0.089  -0.078
##   0.017   0.013
##
## Variances:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .Y        0.855   0.106   8.030   0.000   0.618   1.027
## .M1      0.970   0.118   8.221   0.000   0.731   1.181
## .M2      1.415   0.193   7.328   0.000   1.014   1.792
## Std.lv Std.all
##   0.855   0.465
##   0.970   0.748
##   1.415   0.886
##
## R-Square:
##             Estimate
## Y          0.535
## M1         0.252
## M2         0.114
##

```

```

## Defined Parameters:
##                               Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## indirect1                  0.233   0.069   3.381   0.001   0.107    0.381
## indirect2                 -0.283   0.090  -3.159   0.002  -0.466   -0.100
## contrast                   0.516   0.103   5.007   0.000   0.329    0.725
## total_indircts              -0.051   0.122  -0.415   0.678  -0.299    0.194
## total_c                     -0.021   0.123  -0.167   0.868  -0.277    0.215
## direct                      0.030   0.100   0.301   0.764  -0.176    0.214
## Std.lv   Std.all
## 0.233   0.192
## -0.283  -0.234
## 0.516   0.426
## -0.051  -0.042
## -0.021  -0.017
## 0.030   0.025

```

pfit\_ParEsts

	lhs	op	rhs	label	est	se	
## 1	Y	~	M1	b1	0.456	0.111	
## 2	Y	~	M2	b2	0.743	0.074	
## 3	Y	~	X	c_p	0.030	0.100	
## 4	M1	~	X	a1	0.510	0.079	
## 5	M2	~	X	a2	-0.381	0.121	
## 6	Y	~~	Y		0.855	0.106	
## 7	M1	~~	M1		0.970	0.118	
## 8	M2	~~	M2		1.415	0.193	
## 9	X	~~	X		1.253	0.000	
## 10	Y	~1			0.113	0.092	
## 11	M1	~1			-0.089	0.097	
## 12	M2	~1			0.017	0.120	
## 13	X	~1			0.009	0.000	
## 14	indirect1 :=		a1*b1	indirect1	0.233	0.069	
## 15	indirect2 :=		a2*b2	indirect2	-0.283	0.090	
## 16	contrast :=	indirect1-indirect2		contrast	0.516	0.103	
## 17	total_indirects :=	indirect1+indirect2	total_indirects	-0.051	0.122		
## 18	total_c := c_p+(indirect1)+(indirect2)		total_c	-0.021	0.123		
## 19	direct :=	c_p	direct	0.030	0.100		
	z	pvalue	ci.lower	ci.upper	std.lv	std.all	std.nox
## 1	4.123	0.000	0.249	0.679	0.456	0.383	0.383
## 2	10.095	0.000	0.612	0.906	0.743	0.693	0.693
## 3	0.301	0.764	-0.174	0.215	0.030	0.025	0.022
## 4	6.480	0.000	0.355	0.664	0.510	0.502	0.448
## 5	-3.152	0.002	-0.609	-0.124	-0.381	-0.338	-0.302
## 6	8.030	0.000	0.676	1.102	0.855	0.465	0.465
## 7	8.221	0.000	0.773	1.222	0.970	0.748	0.748
## 8	7.328	0.000	1.097	1.872	1.415	0.886	0.886

```

## 9      NA      NA   1.253   1.253   1.253   1.000   1.253
## 10    1.224  0.221  -0.068   0.289   0.113   0.083   0.083
## 11   -0.913  0.361  -0.278   0.099  -0.089  -0.078  -0.078
## 12    0.140  0.888  -0.231   0.237   0.017   0.013   0.013
## 13      NA      NA   0.009   0.009   0.009   0.008   0.009
## 14    3.381  0.001   0.113   0.390   0.233   0.192   0.172
## 15   -3.159  0.002  -0.459  -0.097  -0.283  -0.234  -0.209
## 16    5.007  0.000   0.316   0.711   0.516   0.426   0.380
## 17   -0.415  0.678  -0.292   0.206  -0.051  -0.042  -0.037
## 18   -0.167  0.868  -0.262   0.240  -0.021  -0.017  -0.015
## 19    0.301  0.764  -0.174   0.215   0.030   0.025   0.022

```

#### 6.4.1.3 A note on indirect effects and confidence intervals

Before we move onto interpretation, I want to stop and look at both  $p$  values and confidence intervals. Especially with Hayes [2022b] PROCESS macro, there is a great deal of emphasis on the use of bootstrapped confidence intervals to determine the statistical significance of the indirect effects. In fact, PROCESS output has (at least historically) not provided  $p$  values with the indirect effects. This is because, especially in the ordinary least squares context, bias-corrected bootstrapped confidence intervals are more powerful (i.e., they are more likely to support a statistically significant result) than  $p$  values.

An excellent demonstration of this phenomena was provided by Mallinckrodt et al. [2006] where they compared confidence intervals produced by the normal theory method to those that are bias corrected. The bias corrected intervals were more powerful to determining if there were statistically significant indirect effects.

The method we have specified in *lavaan* produced bias-corrected confidence intervals. The  $p$  values and corresponding confidence intervals should be consistent with each other. That is, if  $p < .05$ , then the CI95s should not pass through zero. Of course we can always check to be certain this is true. For this reason, I will report  $p$  values in my results. There are reviewers, though, who may prefer that you report CI95s (or both).

#### 6.4.1.4 Figures and Tables

To assist in table preparation, it is possible to export the results to a .csv file that can be manipulated in Excel, Microsoft Word, or other program to prepare an APA style table.

```
write.csv(pfit_ParEsts, file = "pfit_ParEsts.csv")
```

We can use the package [tidySEM](#) to create a figure that includes the values on the path.

Here's what the base package gets us

```
# only worked when I used the library to turn on all these pkgs
library(lavaan)
```

```
## This is lavaan 0.6-16
## lavaan is FREE software! Please report any bugs.
```

```
library(dplyr)

## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

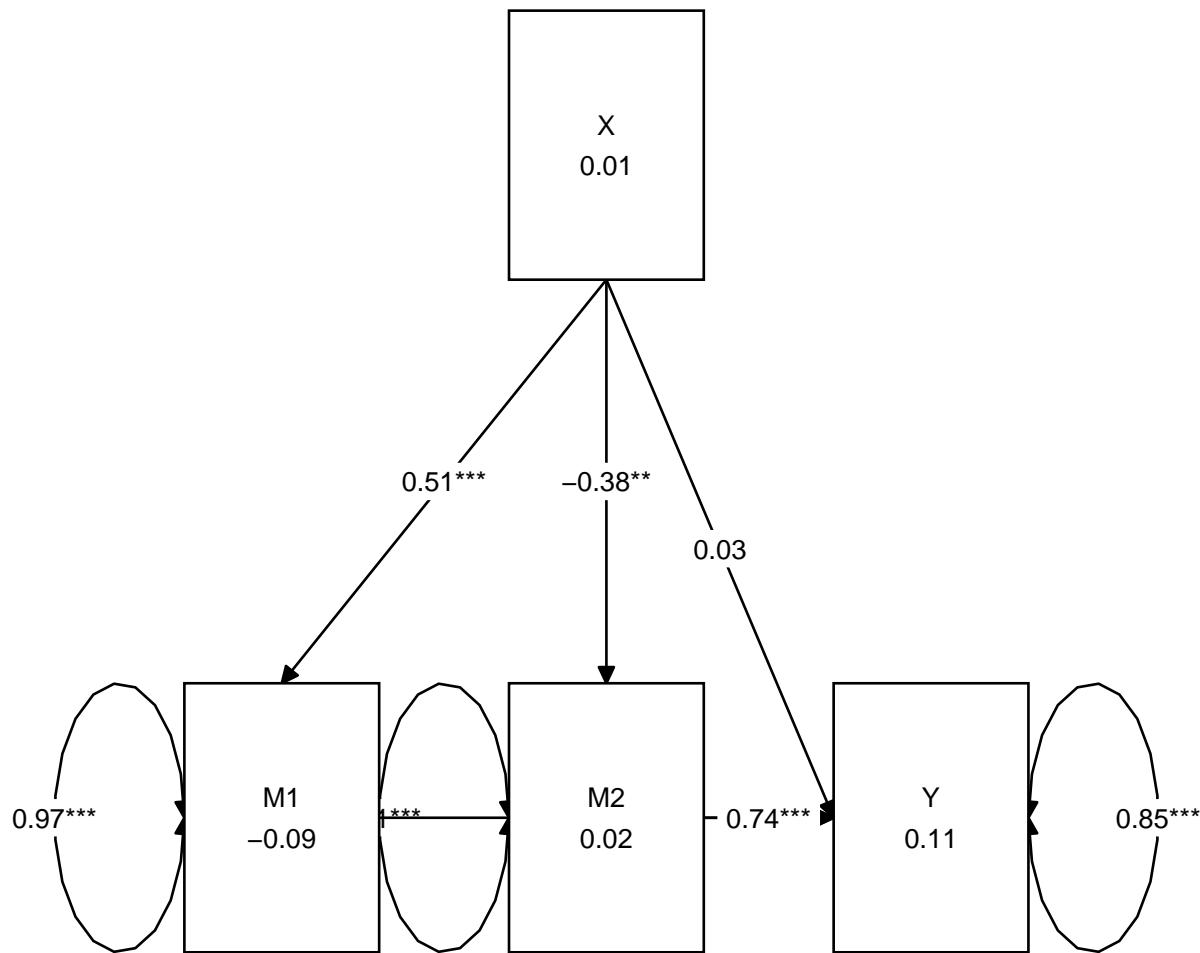
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidySEM)

## Loading required package: OpenMx

## Registered S3 method overwritten by 'tidySEM':
##   method      from
##   predict.MxModel  OpenMx
```

```
tidySEM::graph_sem(model = parallel_fit)
```



We can create model that communicates more intuitively with a little tinkering. First, let's retrieve the current "map" of the layout.

```
tidySEM::get_layout(parallel_fit)
```

```
##      [,1] [,2] [,3]
## [1,] NA   "X"  NA
## [2,] "M1" "M2" "Y"
## attr(),"class")
## [1] "layout_matrix" "matrix"       "array"
```

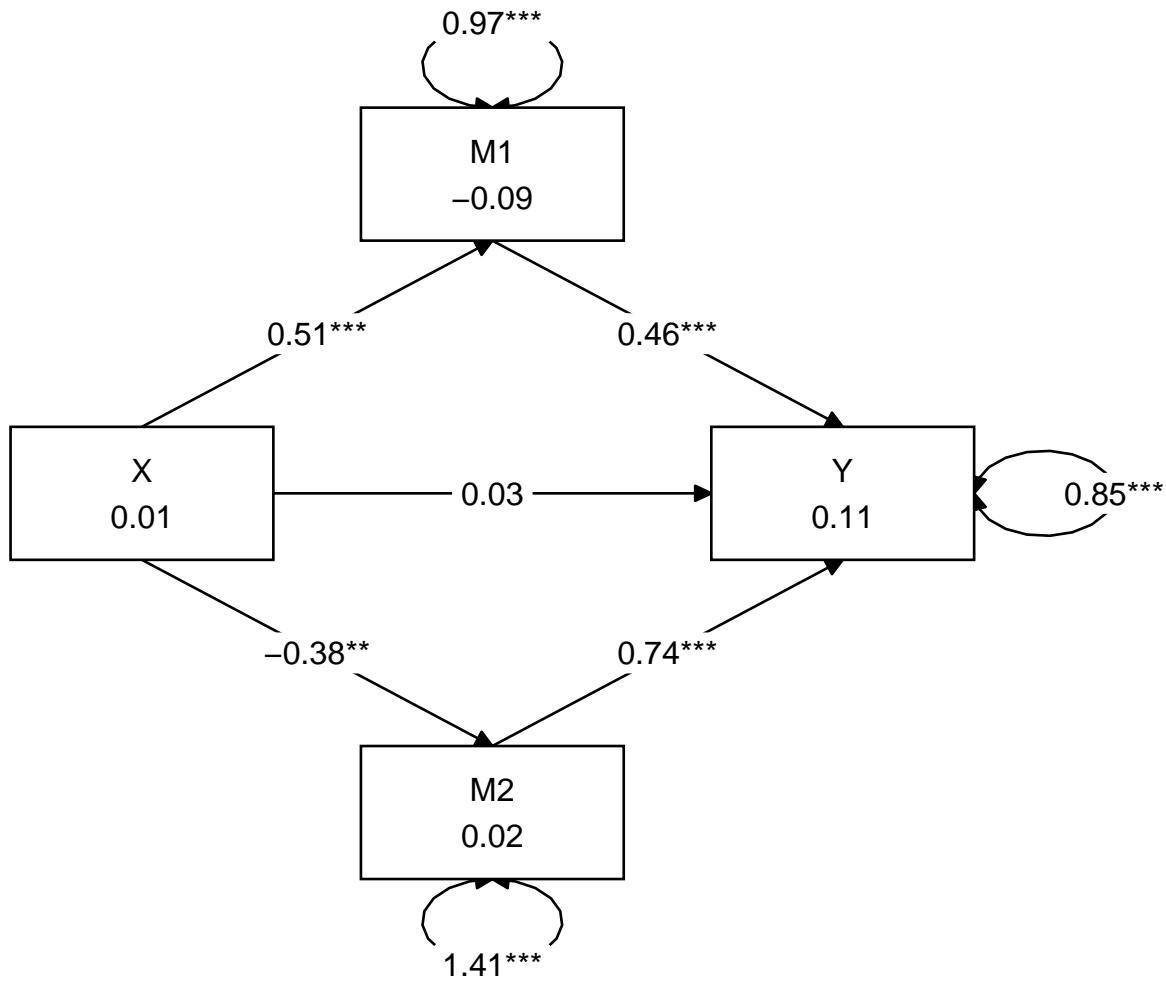
To create the figure I showed at the beginning of the chapter, we will want three rows and three columns.

```
parallel_map <- tidySEM::get_layout("", "M1", "", "X", "", "Y", "", "M2",
 "", rows = 3)
parallel_map
```

```
##      [,1] [,2] [,3]
## [1,] ""   "M1"  ""
## [2,] "X"  ""   "Y"
## [3,] ""   "M2"  ""
## attr(),"class")
## [1] "layout_matrix" "matrix"       "array"
```

We can update our figure by supplying this new map and adjusting the object and text sizes.

```
tidySEM::graph_sem(parallel_fit, layout = parallel_map, rect_width = 1.5,
rect_height = 1.25, spacing_x = 2, spacing_y = 3, text_size = 4.5)
```



There are a number of ways to tabalize the data. You might be surprised to learn that a number of articles that analyze mediating effects focus their presentation on those values and not the traditional intercepts and B weights. This is the approach I have taken in this chapter.

**Table 1**

---

 Model Coefficients Assessing M1 and M2 as Parallel Mediators Between X and Y
 

---

Predictor	B	SE <sub>B</sub>	p	R <sup>2</sup>
M1				.25
Constant	-0.089	0.097	0.360	
X ( $a_1$ )	0.510	0.076	0.000	
M2				.11
Constant	0.017	0.126	0.894	
X ( $a_2$ )	-0.381	0.117	0.001	
DV				.54
Constant	0.113	0.097	0.243	
M1 ( $b_1$ )	0.456	0.113	<0.001	
M2 ( $b_2$ )	0.743	0.074	<0.001	
X ( $c'$ )	0.030	0.098	0.757	
Summary of Effects	B	SE <sub>B</sub>	p	95% CI
Total	-0.021	0.120	0.865	-0.251, 0.214
Indirect 1 ( $a_1 * a_2$ )	0.233	0.070	0.001	0.116, 0.394
Indirect 2 ( $b_1 * b_2$ )	-0.283	0.086	0.001	-0.455, -0.106
Total indirects	-0.051	0.121	0.676	-0.280, 0.187
Contrast (Ind1 - Ind2)	0.516	0.100	0.000	0.324, 0.725

---

Note. The significance of the indirect effects was calculated with bootstrapped, bias-corrected, confidence intervals (.95).

---

#### 6.4.1.5 APA Style Writeup

You may notice that my write-up includes almost no statistical output. This is consistent with APA style that avoids redundancy in text and table. When I want to emphasize a specific result, I may duplicate some output in the text.

A model of parallel multiple mediation was analyzed examining the degree to which importance of M1 and M2 mediated the relation of X on Y. Hayes [2022b] recommended this strategy over simple mediation models because it allows for all mediators to be

examined, simultaneously. The resultant direct and indirect values for each path account for other mediation paths. Using the *lavaan* (*v. 0.6-16*) package in R, coefficients for specific indirect, total indirect, direct, and total were computed. Path coefficients refer to regression weights, or slopes, of the expected changes in the dependent variable given a unit change in the independent variables.

Results (depicted in Figure 1 and presented in Table 1) suggest that 54% of the variance in Y is accounted for by the model. Neither the total nor direct effect of X on Y were statistically significant. In contrast, both indirect effects were statistically significant. A pairwise comparison of the specific indirect effects indicated that the strength of the effects were statistically significantly different from each other. In summary, the effect of X on Y is mediated through M1 and M2, with a stronger influence through M2.

You may notice this write-up included only one statistic. I offered this as an example of avoiding redundancy in text and table. When tables and figures convey maximal information, the results section may be used to describe the patterns – including numbers when they reduce the cognitive load for the readers and reviewers.

Let's turn now to the research vignette and work an example with simulated data from that example. Because the research vignette use an entirely new set of output I will either restart R or clear my environment so that there are a few less objects “in the way.”

#### 6.4.2 Research Vignette

The research vignette comes from the Lewis, Williams, Peppers, and Gadson’s [2017] study titled, “Applying Intersectionality to Explore the Relations Between Gendered Racism and Health Among Black Women.” The study was published in the Journal of Counseling Psychology. Participants were 231 Black women who completed an online survey.

Variables used in the study included:

- **GRMS:** Gendered Racial Microaggressions Scale [Lewis and Neville, 2015] is a 26-item scale that assesses the frequency of nonverbal, verbal, and behavioral negative racial and gender slights experienced by Black women. Scaling is along six points ranging from 0 (*never*) to 5 (*once a week or more*). Higher scores indicate a greater frequency of gendered racial microaggressions. An example item is, “Someone has tried to ‘put me in my place.’”
- **MntlHlth** and **PhysHlth:** Short Form Health Survey - Version 2 [Ware et al., 1995] is a 12-item scale used to report self-reported mental (six items) and physical health (six items). Although the article did not specify, when this scale is used in other contexts [e.g., Kim et al., 2017], a 6-point scale has been reported. Higher scores indicate higher mental health (e.g., little or no psychological distress) and physical health (e.g., little or no reported symptoms in physical functioning). An example of an item assessing mental health was, “How much of the time during the last 4 weeks have you felt calm and peaceful?”; an example of a physical health item was, “During the past 4 weeks, how much did pain interfere with your normal work?”

- **Sprtlty, SocSup, Engmgt, and DisEngmt** are four subscales from the Brief Coping with Problems Experienced Inventory [Carver, 1997]. The 28 items on this scale are presented on a 4-point scale ranging from 1 (*I usually do not do this at all*) to 4(*I usually do this a lot*). Higher scores indicate a respondents' tendency to engage in a particular strategy. Instructions were modified to ask how the female participants responded to recent experiences of racism and sexism as Black women. The four subscales included spirituality (religion, acceptance, planning), interconnectedness/social support (vent emotions, emotional support, instrumental social support), problem-oriented/engagement coping (active coping, humor, positive reinterpretation/positive reframing), and disengagement coping (behavioral disengagement, substance abuse, denial, self-blame, self-distraction).
- **GRICentlty:** The Multidimensional Inventory of Black Identity Centrality subscale [Sellers et al.] was modified to measure the intersection of racial and gender identity centrality. The scale included 10 items scaled from 1 (*strongly disagree*) to 7 (*strongly agree*). An example item was, “Being a *Black woman* is important to my self-image.” Higher scores indicated higher levels of gendered racial identity centrality.

#### 6.4.2.1 Data Simulation

The *lavaan::simulateData* function was used. If you have taken psychometrics, you may recognize the code as one that creates latent variables from item-level data. In trying to be as authentic as possible, we retrieved factor loadings from psychometrically oriented articles that evaluated the measures [Nadal, 2011, Veit and Ware, 1983]. For all others we specified a factor loading of 0.80. We then approximated the *measurement model* by specifying the correlations between the latent variable. We sourced these from the correlation matrix from the research vignette [Lewis et al., 2017]. The process created data with multiple decimals and values that exceeded the boundaries of the variables. For example, in all scales there were negative values. Therefore, the final element of the simulation was a linear transformation that rescaled the variables back to the range described in the journal article and rounding the values to integer (i.e., with no decimal places).

#Entering the intercorrelations, means, and standard deviations from the journal article

```
Lewis_generating_model <- '
  ##measurement model
  GRMS  =~ .69*Ob1 + .69*Ob2 + .60*Ob3 + .59*Ob4 + .55*Ob5 + .55*Ob6 + .54*Ob7 + .50*Ob8
  MntlHlth  =~ .8*MH1 + .8*MH2 + .8*MH3 + .8*MH4 + .8*MH5 + .8*MH6
  PhysHlth  =~ .8*PhH1 + .8*PhH2 + .8*PhH3 + .8*PhH4 + .8*PhH5 + .8*PhH6
  Spirituality  =~ .8*Spirit1 + .8*Spirit2
  SocSupport  =~ .8*SocS1 + .8*SocS2
  Engagement  =~ .8*Eng1 + .8*Eng2
  Disengagement  =~ .8*dEng1 + .8*dEng2
  GRIC  =~ .8*Cntrlty1 + .8*Cntrlty2 + .8*Cntrlty3 + .8*Cntrlty4 + .8*Cntrlty5 + .8*Cntrlty6

  # Means
  GRMS ~ 1.99*1
  Spirituality ~2.82*1
  SocSupport ~ 2.48*1
```

```

Engagement ~ 2.32*1
Disengagement ~ 1.75*1
GRIC ~ 5.71*1
MntlHlth ~3.56*1 #Lewis et al used sums instead of means, I recast as means to facilitate comparison
PhysHlth ~ 3.51*1 #Lewis et al used sums instead of means, I recast as means to facilitate comparison

# Correlations (ha!)
GRMS ~ 0.20*Spirituality
GRMS ~ 0.28*SocSupport
GRMS ~ 0.30*Engagement
GRMS ~ 0.41*Disengagement
GRMS ~ 0.19*GRIC
GRMS ~ -0.32*MntlHlth
GRMS ~ -0.18*PhysHlth

Spirituality ~ 0.49*SocSupport
Spirituality ~ 0.57*Engagement
Spirituality ~ 0.22*Disengagement
Spirituality ~ 0.12*GRIC
Spirituality ~ -0.06*MntlHlth
Spirituality ~ -0.13*PhysHlth

SocSupport ~ 0.46*Engagement
SocSupport ~ 0.26*Disengagement
SocSupport ~ 0.38*GRIC
SocSupport ~ -0.18*MntlHlth
SocSupport ~ -0.08*PhysHlth

Engagement ~ 0.37*Disengagement
Engagement ~ 0.08*GRIC
Engagement ~ -0.14*MntlHlth
Engagement ~ -0.06*PhysHlth

Disengagement ~ 0.05*GRIC
Disengagement ~ -0.54*MntlHlth
Disengagement ~ -0.28*PhysHlth

GRIC ~ -0.10*MntlHlth
GRIC ~ 0.14*PhysHlth

MntlHlth ~ 0.47*PhysHlth
'

set.seed(230925)
dfLewis <- lavaan::simulateData(model = Lewis_generating_model,
                                model.type = "sem",
                                meanstructure = T,

```

```

        sample.nobs=231,
        standardized=FALSE)

#used to retrieve column indices used in the rescaling script below
#col_index <- as.data.frame(colnames(dfLewis))

for(i in 1:ncol(dfLewis)){ # for loop to go through each column of the dataframe
  if(i >= 1 & i <= 25){ # apply only to GRMS variables
    dfLewis[,i] <- scales::rescale(dfLewis[,i], c(0, 5))
  }
  if(i >= 26 & i <= 37){ # apply only to mental and physical health variables
    dfLewis[,i] <- scales::rescale(dfLewis[,i], c(0, 6))
  }
  if(i >= 38 & i <= 45){ # apply only to coping variables
    dfLewis[,i] <- scales::rescale(dfLewis[,i], c(1, 4))
  }
  if(i >= 46 & i <= 55){ # apply only to GRIC variables
    dfLewis[,i] <- scales::rescale(dfLewis[,i], c(1, 7))
  }
}

#rounding to integers so that the data resembles that which was collected
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## vforcats 1.0.0   vstringr 1.5.0
## vlubridate 1.9.2   vtibble 3.2.1
## vpurrr 1.0.1   vtidy 1.3.0
## vreadr 2.1.4
## -- Conflicts ----- tidyverse_conflicts() --
## xdplyr::filter() masks stats::filter()
## xdplyr::lag() masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco

dfLewis <- dfLewis %>% round(0)

#quick check of my work
#psych::describe(dfLewis)

```

The script below allows you to store the simulated data as a file on your computer. This is optional – the entire lesson can be worked with the simulated data.

If you prefer the .rds format, use this script (remove the hashtags). The .rds format has the advantage of preserving any formatting of variables. A disadvantage is that you cannot open these files outside of the R environment.

Script to save the data to your computer as an .rds file.

```
#saveRDS(dfLewis, 'dfLewis.rds')
```

Once saved, you could clean your environment and bring the data back in from its .csv format.

```
#dfLewis<- readRDS('dfLewis.rds')
```

If you prefer the .csv format (think “Excel lite”) use this script (remove the hashtags). An advantage of the .csv format is that you can open the data outside of the R environment. A disadvantage is that it may not retain any formatting of variables

Script to save the data to your computer as a .csv file.

```
# write.table(dfLewis, file = 'dfLewis.csv', sep = ',',
# col.names=TRUE, row.names=FALSE)
```

Once saved, you could clean your environment and bring the data back in from its .csv format.

```
#dfLewis<- read.csv ('dfLewis.csv', header = TRUE)
```

### 6.4.3 Scrubbing, Scoring, and Data Diagnostics

Because the focus of this lesson is on complex mediation, we have used simulated data. If this were real, raw, data, it would be important to [scrub](#), [score](#), and conduct [data diagnostics](#) to evaluate the suitability of the data for the proposed analyses.

Because we are working with item level data we do need to score the scales used in the researcher’s model. Because we are using simulated data and the authors already reverse coded any such items, we will omit that step.

As described in the [Scoring](#) chapter, we calculate mean scores of these variables by first creating concatenated lists of variable names. Next we apply the `sjstats::mean_n` function to obtain mean scores when a given percentage (we’ll specify 80%) of variables are non-missing. Functionally, this would require the two-item variables (e.g., engagement coping and disengagement coping) to have non-missingness. We simulated a set of data that does not have missingness, none-the-less, this specification is useful in real-world settings.

Note that I am only scoring the variables used in the models demonstrated in this lesson. The remaining variables are available as practice options.

```
GRMS_vars <- c("Ob1", "Ob2", "Ob3", "Ob4", "Ob5", "Ob6", "Ob7", "Ob8",
               "Ob9", "Ob10", "Ma1", "Ma2", "Ma3", "Ma4", "Ma5", "Ma6", "Ma7", "St1",
               "St2", "St3", "St4", "St5", "An1", "An2", "An3")
Eng_vars <- c("Eng1", "Eng2")
dEng_vars <- c("dEng1", "dEng2")
MntlHlth_vars <- c("MH1", "MH2", "MH3", "MH4", "MH5", "MH6")

dfLewis$GRMS <- sjstats::mean_n(dfLewis[, GRMS_vars], 0.8)
```

```
dfLewis$Engmt <- sjstats::mean_n(dfLewis[, Eng_vars], 0.8)
dfLewis$DisEngmt <- sjstats::mean_n(dfLewis[, dEng_vars], 0.8)
dfLewis$MntlHlth <- sjstats::mean_n(dfLewis[, MntlHlth_vars], 0.8)
```

Now that we have scored our data, let's trim the variables to just those we need.

```
Lewis_df <- dplyr::select(dfLewis, GRMS, Engmt, DisEngmt, MntlHlth)
```

Let's check a table of means, standard deviations, and correlations to see if they align with the published article.

```
Lewis_table <- apaTables::apa.cor.table(Lewis_df, table.number = 1, show.sig.stars = TRUE,
                                         landscape = TRUE, filename = "Lewis_Corr.doc")
print(Lewis_table)
```

```
##
##
## Table 1
##
## Means, standard deviations, and correlations with confidence intervals
##
##
##      Variable    M     SD   1          2          3
## 1. GRMS    2.56 0.72
##
## 2. Engmt    2.48 0.53 .52**    [.42, .61]
##
## 3. DisEngmt 2.48 0.52 .53**    .32**    [.43, .62]  [.20, .43]
##
## 4. MntlHlth 3.16 0.81 -.56**   -.23**   -.48**  [-.64, -.47] [-.35, -.11] [-.57, -.37]
##
##
## Note. M and SD are used to represent mean and standard deviation, respectively.
## Values in square brackets indicate the 95% confidence interval.
## The confidence interval is a plausible range of population correlations
## that could have caused the sample correlation (Cumming, 2014).
## * indicates p < .05. ** indicates p < .01.
##
```

While they are not exact, they approximate the magnitude and patterns in the correlation matrix in the article [Lewis et al., 2017].

### 6.4.3.1 Specifying the *lavaan* model

The Lewis et al. article [2017] reports four mediation analyses, each repeated for mental and physical outcomes. Thus, their write-up reports eight simple mediation models. Graphically, their analyses were efficiently presented in a figure that looked (to me) like parallel mediation. Correspondingly, it made sense to me that we could try this in our research vignette. In the upcoming chapter on conditional process analysis, we will work the moderated mediation that was a primary focus of their research.

Below is the model we will work. Specifically, we will evaluate whether gendered racial microaggressions impact mental health separately, thorough mediated paths of engagement and disengagement. We will also be able to see if the strength of those mediated paths are statistically, significantly, different from each other.

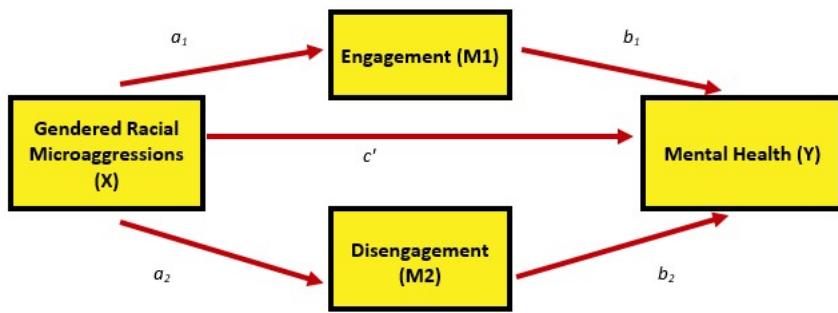


Figure 6.2: An image of the parallel mediation we will work

We can use the guidelines above to specify our model and then request summaries of the fit indices and parameter estimates.

```

parallel_Lewis <- "
  MntlHlth ~ b1*Engmt + b2*DisEngmt + c_p*GRMS
  Engmt ~ a1*GRMS
  DisEngmt ~ a2*GRMS

  indirect1 := a1 * b1
  indirect2 := a2 * b2
  contrast := indirect1 - indirect2
  total_indirects := indirect1 + indirect2
  total_c := c_p + (indirect1) + (indirect2)
  direct := c_p
"
para_Lewis_fit <- lavaan::sem(parallel_Lewis, data = Lewis_df, se = "bootstrap",
  bootstrap = 1000, missing = "fiml") #holds the 'whole' result
pLewis_sum <- lavaan::summary(para_Lewis_fit, standardized = TRUE, rsq = T,
  fit = TRUE, ci = TRUE) #today, we really only need the R-squared from here
pLewis_ParEsts <- lavaan::parameterEstimates(para_Lewis_fit, boot.ci.type = "bca.simple",
  standardized = TRUE) #provides our estimates, se, p values for all the elements we specif
  
```

```
pLewis_sum  
pLewis_ParEsts
```

#### 6.4.3.2 Table and Figure

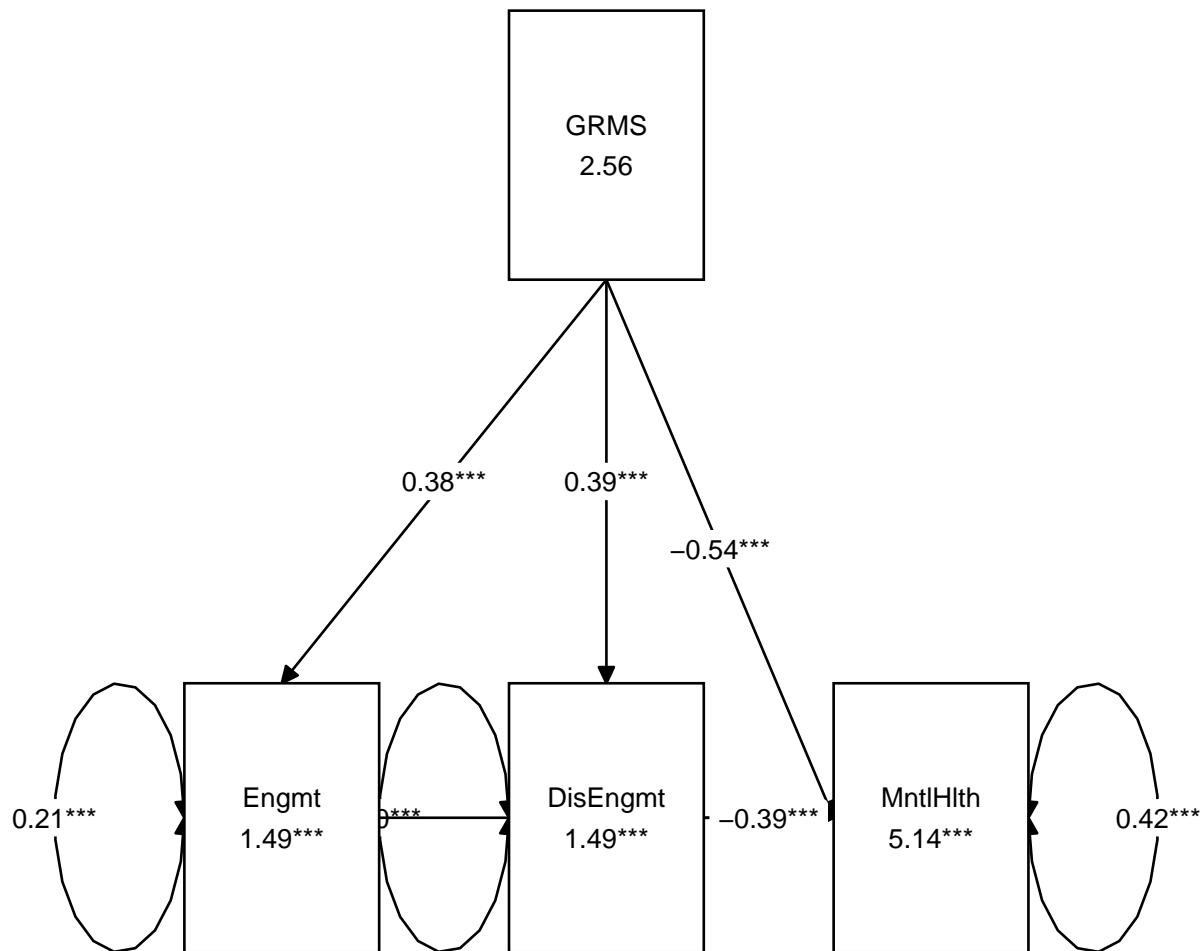
To assist in table preparation, it is possible to export the results to a .csv file that can be manipulated in Excel, Microsoft Word, or other program to prepare an APA style table.

```
write.csv(pLewis_ParEsts, file = "pLewis_ParEsts.csv")
```

We can use the package [tidySEM](#) to create a figure that includes the values on the path.

Here's what the base package gets us

```
# only worked when I used the library to turn on all these pkgs  
library(lavaan)  
library(dplyr)  
library(ggplot2)  
library(tidySEM)  
tidySEM::graph_sem(model = para_Lewis_fit)
```



We can create model that communicates more intuitively with a little tinkering. First, let's retrieve the current "map" of the layout.

```
tidySEM::get_layout(para_Lewis_fit)
```

```
##      [,1]     [,2]     [,3]
## [1,] NA     "GRMS"    NA
## [2,] "Engmt" "DisEngmt" "MntlHlth"
## attr(,"class")
## [1] "layout_matrix" "matrix"       "array"
```

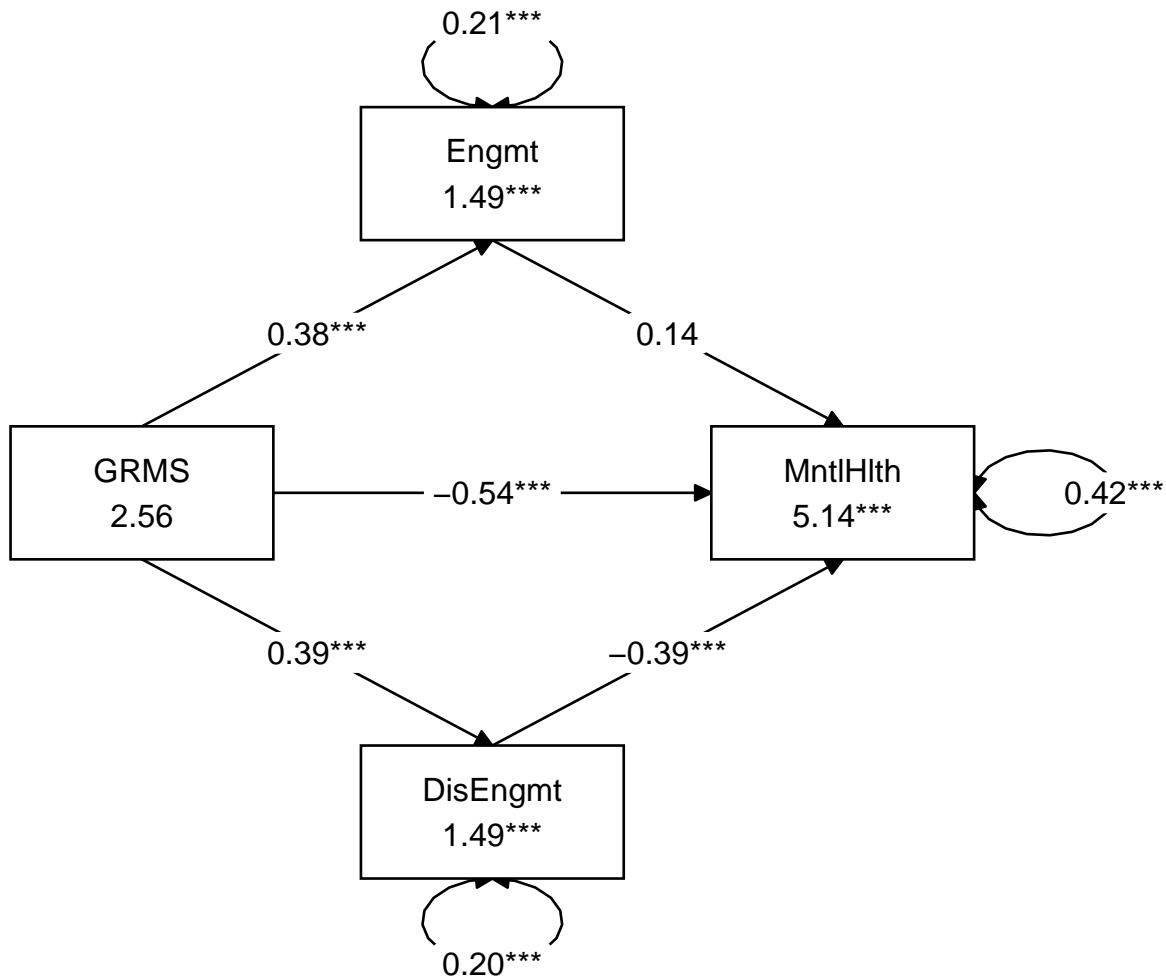
To create the figure I showed at the beginning of the chapter, we will want three rows and three columns.

```
pLewis_map <- tidySEM::get_layout("", "Engmt", "", "GRMS", "", "MntlHlth",
    "", "DisEngmt", "", rows = 3)
pLewis_map
```

```
##      [,1]   [,2]      [,3]
## [1,] ""    "Engmt"    ""
## [2,] "GRMS" ""    "MntlHlth"
## [3,] ""    "DisEngmt" ""
## attr(,"class")
## [1] "layout_matrix" "matrix"      "array"
```

We can update our figure by supplying this new map and adjusting the object and text sizes.

```
tidySEM::graph_sem(para_Lewis_fit, layout = pLewis_map, rect_width = 1.5,
rect_height = 1.25, spacing_x = 2, spacing_y = 3, text_size = 4.5)
```



Now let's make a table.

**Table 2**

---

 Model Coefficients Assessing Engagement and Disengagement Coping as Parallel Mediators  
 Between Predicting Mental Health from Gendered Racial Microaggressions
 

---

Predictor	<i>B</i>	<i>SE<sub>B</sub></i>	<i>p</i>	<i>R<sup>2</sup></i>
Engagement coping (M1)				.27
Constant	1.494	0.109	<0.001	
GRMS ( $a_1$ )	0.384	0.042	<0.001	
Disengagement coping (M2)				.28
Constant	1.490	0.113	<0.001	
GRMS ( $a_2$ )	0.386	0.043	<0.001	
Mental Health (DV)				.37
Constant	5.141	0.239	<0.001	
Engagement ( $b_1$ )	0.144	0.090	0.109	
Disengagement ( $b_2$ )	-0.391	0.089	<0.001	
GRMS ( $c'$ )	-0.535	0.076	<0.001	
Summary of Effects	<i>B</i>	<i>SE<sub>B</sub></i>	<i>p</i>	95% CI
Total	-0.631	0.060	<0.001	-0.748, -0.507
Indirect 1 ( $a_1 * a_2$ )	0.055	0.036	0.121	-0.009, 0.126
Indirect 2 ( $b_1 * b_2$ )	-0.151	0.039	<0.001	-0.230, -0.079
Total indirects	-0.096	0.054	0.075	-0.206, 0.008
Contrast (Ind1 - Ind2)	0.206	0.052	<0.001	0.112, 0.316

---

Note. GRMS = gendered racial microaggressions. The significance of the indirect effects was calculated with bootstrapped, bias-corrected, confidence intervals (.95).

---

- The model accounts for 37% of the variance in predicting mental health outcomes.
- The total effect of GRMS on mental health is -0.631 ( $p < 0.001$ ) is negative and statistically significant. That is, gendered racial microaggressions have a statistically significant negative effect on mental health.
- The direct effect of GRMS on mental health is -0.535 ( $p < 0.001$ ); while this is lower than the total effect, it remains statistically significant.

- Using Baron and Kenny's [1986] causal steps logic, the fact that the direct effect does not decrease in a statistically significant manner does not provide helpful, logical support for mediation. According to Hayes [2022b] this difference is not necessary. That is, a statistically significant indirect effect can stand on its own.
- Indirect effect #1 ( $a_1 \times b_1$  or GRMS through engagement coping) is 0.055 ( $p = 0.121, CI_{95}[-0.011, 0.124]$ ) and not statistically significant. Because they can be inconsistent with the  $p$  values, we should always check the confidence intervals to see if they pass through zero. In this case they do.
- Indirect effect #2 ( $a_2 \times b_2$ , or GRMS through disengagement to coping) is -0.151 ( $p < 0.001, CI_{95}[-0.231, -0.082]$ ). The  $p$  value is significant and the 95% confidence interval does not pass through zero. Thus, gendered racial microaggressions lead to greater disengagement ( $a_1$ ). In turn, disengagement has negative effects on mental health ( $b_2$ ).
- The total indirect effect (i.e., sum of all specific indirect effects) ( $-0.096, p = 0.075$ ) is not statistically significant.
- We examine the contrast to see if the indirect effects statistically significantly different from each other:  $B = 0.206, p < 0.001$ . They are. This is not surprising since the path mediated by engagement was not statistically significant but the path mediated by disengagement was statistically significant.

#### 6.4.3.3 APA Style Writeup

Hayes [Hayes, 2022a] provides helpful guidelines for presenting statistical results. Here is a summary of his recommendations.

- Pack as much statistical info as possible into a table(s) or figure(s).
- Use statistics in the text as punctuation; avoid redundancy in text and table.
- Avoid using abbreviations for variables in the text itself; rather focus on the construct names rather than their shorthand
- Avoid focusing on what you hypothesized (e.g., avoid, "Results supported/did not support hypothesis A1") and instead focus on what you found. The reader is more interested in the results, not your forecasts.
- Hayes prefers reporting unstandardized metrics because they map onto the measurement scales used in the study. He believes this is especially important when dichotomous variables are used.
- There is "no harm" in reporting hypothesis tests and CIs for the  $a$  and  $b$  paths, but whether/not these paths are statistically significant does not determine the significance of the indirect effect.
- Be precise with language:
  - OK: X exerts an effect on Y directly and/or indirectly through M.
  - Not OK: the indirect effect of M
- Report direct and indirect effects and their corresponding inferential tests
- Hayes argues that a statistically significant indirect effect is, in fact statistic. He dislikes narration of the Baron and Kenny [1986] process and steps.

Here's my attempt to write up the simulated data from the Lewis et al. [2017] article.

## Method

### Data Analysis

Parallel multiple mediation is appropriate when testing the influence of an independent variable (X) on the dependent variable (Y) directly, as well as indirectly through two or more mediators. A condition of parallel multiple mediation is that no mediator causally influences another [Hayes, 2022b]. Using data simulated from Lewis et al. [2017] we utilized parallel multiple mediation analysis to test the influence of gendered racial microaggressions (X, GRMS) on mental health outcomes (Y, MntlHlth) directly as well as indirectly through the mediators engagement coping (M1, Engmt) and disengaged coping (M2, DisEngmt). Using the *lavaan* (v. 0.6-16) package in R we followed the procedures outlined in Hayes [2022b] by analyzing the strength and significance of four sets of effects: specific indirect, the total indirect, the direct, and total.

## Results

**Preliminary Analyses** Descriptive statistics were computed, and all variables were assessed for skewness and kurtosis. *More narration, here.* A summary of descriptive statistics and a correlation matrix for the study is provided in Table 2. These bivariate relations provide evidence to support the test of mediation analysis.

**Parallel Multiple Mediation Analysis** A model of parallel mediation examined the degree to which engagement and disengagement coping strategies mediated the relation of gendered racial microaggressions on mental health outcomes in Black women. Hayes [2022b] recommended this strategy over simple mediation models because it allows for all mediators to be examined, simultaneously. The resultant direct and indirect values for each path account for other mediation paths. Using the *lavaan* (v. 0.6-17) package in R, coefficients for specific indirect, total indirect, direct, and total were computed. Path coefficients refer to regression weights, or slopes, of the expected changes in the dependent variable given a unit change in the independent variables.

Results (depicted in Figure 2 and presented in Table 3) suggest that 37% of the variance in mental health outcomes is accounted for by the model. The indirect effect predicting mental health from gendered racial microaggressions via engagement coping was not statistically significant  $*B = 0.055, SE = 0.036, p = 0.121, CI95[-0.011, 0.124]$ . Looking at the individual paths we see that  $a_1$  was positive and statistically significant (GRMS leads to increased engagement coping), but the subsequent link,  $b_1$  (engagement to mental health) was not. The indirect effect predicting mental health from gendered racial microaggressions through disengagement to coping was statistically significant  $B = -0.151, SE = 0.039, p < 0.001, CI95[-0.231, -0.082]$ . In this case, gendered racial microaggressions led to greater disengagement coping ( $a_2$ ). In turn, disengagement coping had negative effects on mental health ( $b_2$ ). Curiously, the total indirect effect (i.e., the sum of the specific indirect effects was not statistically significant. It is possible that the positive and negative valences of the indirect effects “cancelled each other out.” A pairwise comparison of the specific indirect effects indicated that the strength of the effects were statistically significantly different from each other. Given that the path through engagement coping was not significant, but the path through disengagement coping was, this statistically significant difference is not surprising.

### Hints for Writing Method/Results Sections

- When you find an article you like, make note of it and put it in a very special folder. In recent years, I have learned to rely on full-text versions stored in my Zotero app.

- Once you know your method (measure, statistic, etc.) begin collecting others articles that are similar to it. To write results sections I will often reference multiple articles.
- When it is time to write have all these resources handy and use them as guides/models.
- Put as much info as possible in the table. Become a table-pro. That is, learn how to merge/split cells, use borders/shading, the decimal tab, and so forth. Don't make the borders disappear until the last thing you do before submitting. This is because you ALWAYS have to update your tables and seeing the borders makes it easier.

## 6.5 Serial Multiple Mediator Model

Recall that one of the conditions of the *parallel mediator model* was that “no mediator causally influences another.”

Regarding these correlated mediators [Hayes, 2022b]:

- Typically, two or more mediators that are causally located between X and Y will be correlated - if for no other reason than that they share a common cause (X).
- Estimating the partial correlation between two mediators after controlling for X is one way to examine whether all of their association is accounted for by this common cause.
- Partial correlation* is the Pearson correlation between the residuals from a model estimating Y from a set of covariates, and the residuals from a model estimating X from the same set of covariates.
- Partial correlations allow the assessment of their association, independent of what they have in common with the covariates that were regressed onto Y and X, separately.
- If two (or more) mediators remain correlated after adjusting for X, then
  - the correlation is *spurious*, they share another (unmodeled) common cause.
  - the remaining association is *epiphenomenal*. That is, a proposed mediator could be related to an outcome not because it causes the outcome, but because it is correlated with another variable that is causally influencing the outcome. This is a noncausal alternative explanation for an association. Also, many things correlated with the cause of Y will also tend to be correlated with X, but it doesn't make all those things cause Y
  - or one mediator causally affects another

The goal of a serial multiple mediator model is to investigate the direct and indirect effects of X on Y while modeling a process in which X causes M<sub>1</sub>, which in turn causes M<sub>2</sub>, and so forth, concluding with Y as the final consequent.

As before, we will calculate:

- Direct effect, c'*: the estimated difference in Y between two cases that differ by one unit on X but who are equal on all mediators in the model.
- Specific indirect effects, a<sub>1</sub>b<sub>1</sub>, a<sub>2</sub>b<sub>2</sub>, a<sub>3</sub>b<sub>3</sub>, etc.*: constructed by multiplying the regression weights corresponding to each step in an indirect pathway; interpreted as the estimated difference in Y between two cases that differ by one unit on X through the causal sequence from X to mediator(s) to Y.
- Total indirect effect of X*: sum of all specific indirect effects

- *Total effect of X*: the total indirect effect of X plus the direct effect of X; can also be estimated by regressing Y from X only.
- *Pairwise comparisons (contrasts) between indirect effects* (i.e., is one indirect effect stronger than another)

### 6.5.1 We stick with the Lewis et al. [2017] example, but modify it.

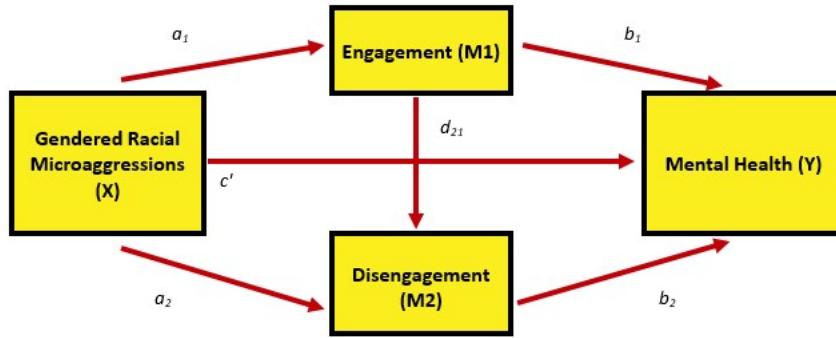


Figure 6.3: An image of the serial mediation we will work

Our parallel multiple mediator model of gendered racial microaggressions on mental health through engagement and disengagement coping strategies assumed no causal association between the mediators. Noting the statistically significant correlation between engagement and disengagement, what if engagement influenced disengagement, which, in turn influenced mental health.

If this is our goal (image), how many direct and indirect effects are contained in this model? Using the same processes as before, let's plan our model:

- We add a path predicting disengagement from engagement, and label it with a  $d_{21}$ 
  - Regarding the notation, it makes sense that we use a  $d$  to designate a new type of path; I don't know why we use a subscript of 21
- We specify a third indirect path that multiplies those 3 paths ( $a_1, d_{21}, b_2$ ) together
- We add a third contrast so that we get all the combinations of indirect comparisons: 1-2, 1-3  
2-3
- We update our total\_indirects calculation to include indirect#3
- We update our total\_c calculation to include indirect#3

### 6.5.2 Specify the *lavaan* model

```

serial_Lewis <- "
  MntlHlth ~ b1*Engmt + b2*DisEngmt + c_p*GRMS
  Engmt ~ a1*GRMS
  DisEngmt ~ a2*GRMS
  DisEngmt ~ d21*Engmt
"
  
```

```

indirect1 := a1 * b1
indirect2 := a2 * b2
indirect3 := a1 * d21 * b2
contrast1 := indirect1 - indirect2
contrast2 := indirect1 - indirect3
contrast3 := indirect2 - indirect3
total_indirects := indirect1 + indirect2 + indirect3
total_c := c_p + indirect1 + indirect2 + indirect3
direct := c_p
"
serial_Lewis_fit <- lavaan::sem(serial_Lewis, data = Lewis_df, se = "bootstrap",
  missing = "fiml", bootstrap = 1000)
sLewis_sum <- lavaan::summary(serial_Lewis_fit, standardized = TRUE, rsq = T,
  fit = TRUE, ci = TRUE)
sLewis_ParEsts <- lavaan::parameterEstimates(serial_Lewis_fit, boot.ci.type = "bca.simple",
  standardized = TRUE)

sLewis_sum
sLewis_ParEsts

```

### 6.5.2.1 Table and Figure

To assist in table preparation, it is possible to export the results to a .csv file that can be manipulated in Excel, Microsoft Word, or other program to prepare an APA style table.

```
write.csv(sLewis_ParEsts, file = "sLewis_ParEsts.csv")
```

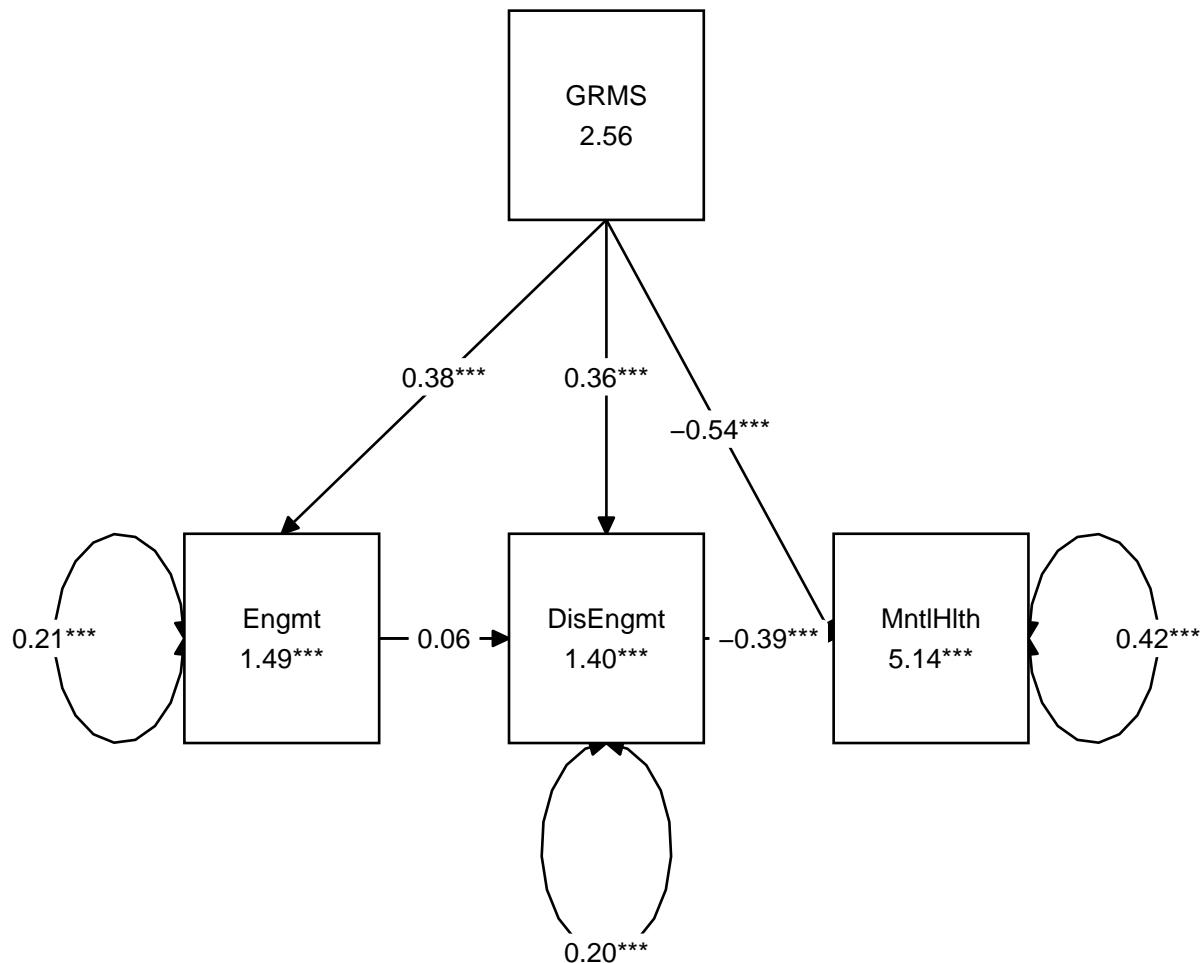
We can use the package [tidySEM](#) to create a figure that includes the values on the path.

Here's what the base package gets us

```

# only worked when I used the library to turn on all these pkgs
library(lavaan)
library(dplyr)
library(ggplot2)
library(tidySEM)
tidySEM::graph_sem(model = serial_Lewis_fit)

```



We can create model that communicates more intuitively with a little tinkering. First, let's retrieve the current “map” of the layout.

```
tidySEM::get_layout(serial_Lewis_fit)
```

```
##      [,1]     [,2]     [,3]
## [1,] NA     "GRMS"    NA
## [2,] "Engmt" "DisEngmt" "MntlHlth"
## attr(,"class")
## [1] "layout_matrix" "matrix"       "array"
```

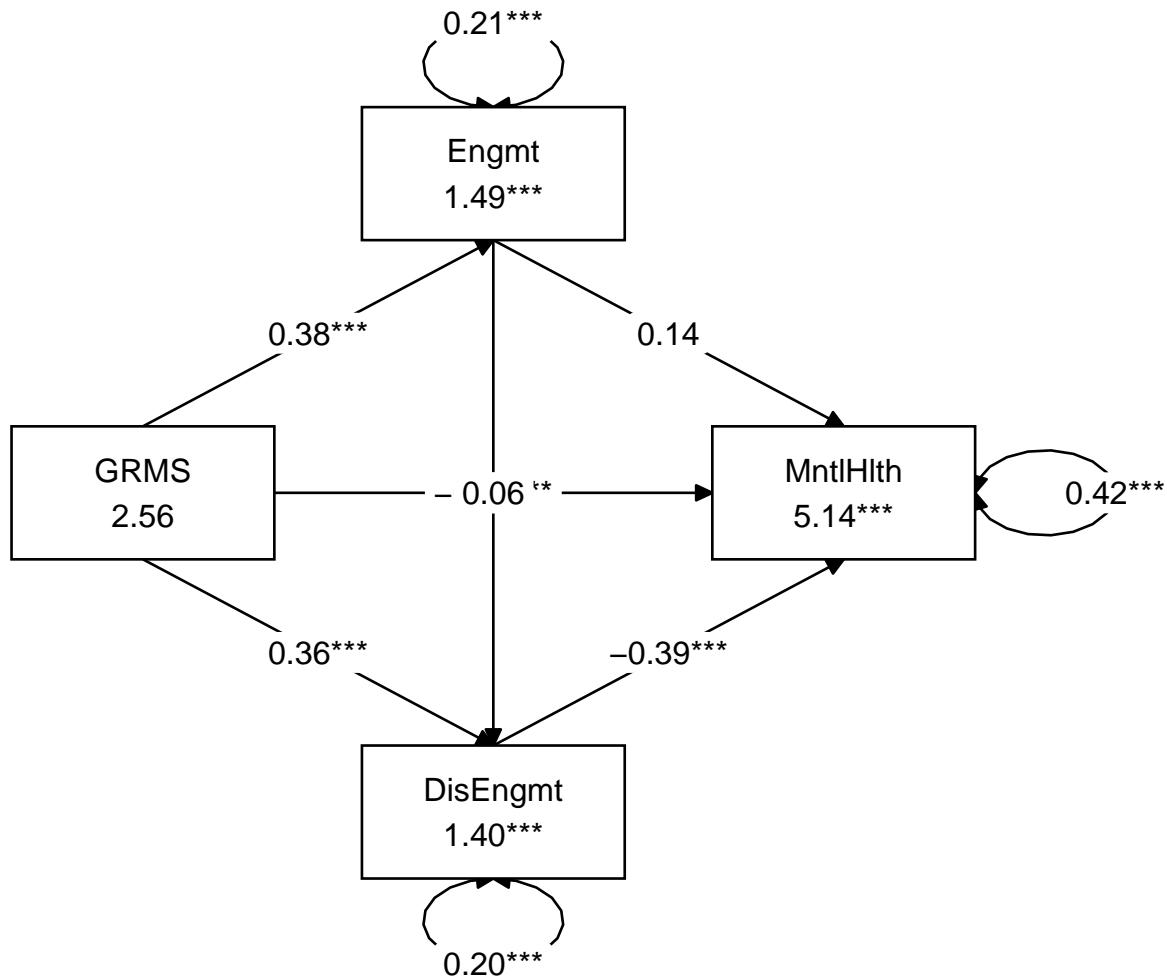
To create the figure I showed at the beginning of the chapter, we will want three rows and three columns.

```
sLewis_map <- tidySEM::get_layout("", "Engmt", "", "GRMS", "", "MntlHlth",
 "", "DisEngmt", "", rows = 3)
sLewis_map
```

```
##      [,1]   [,2]      [,3]
## [1,] ""    "Engmt"    ""
## [2,] "GRMS" ""    "MntlHlth"
## [3,] ""    "DisEngmt" ""
## attr(,"class")
## [1] "layout_matrix" "matrix"      "array"
```

We can update our figure by supplying this new map and adjusting the object and text sizes.

```
tidySEM::graph_sem(serial_Lewis_fit, layout = sLewis_map, rect_width = 1.5,
rect_height = 1.25, spacing_x = 2, spacing_y = 3, text_size = 4.5)
```



Now let's make a table.

**Table 4**

---

 Model Coefficients Assessing Engagement and Disengagement Coping in a Model of Serial Mediation Predicting Mental Health from Gendered Racial Microaggressions
 

---

Predictor	<i>B</i>	<i>SE<sub>B</sub></i>	<i>p</i>	<i>R</i> <sup>2</sup>
Engagement coping (M1)				.27
Constant	1.494	0.112	<0.001	
GRMS ( $a_1$ )	0.384	0.042	<0.001	
Disengagement coping (M2)				.29
Constant	1.400	0.133	<0.001	
GRMS ( $a_2$ )	0.363	0.048	<0.001	
Engagement ( $d_{21}$ )	0.061	0.061	0.321	
Mental Health (DV)				.37
Constant	5.141	0.230	<0.001	
Engagement ( $b_1$ )	0.144	0.089	0.107	
Disengagement ( $b_2$ )	-0.391	0.090	<0.001	
GRMS ( $c'$ )	-0.535	0.077	<0.001	
Effects	<i>B</i>	<i>SE<sub>B</sub></i>	<i>p</i>	95% CI
Total effect	-0.631	0.059	0.000	-0.735, -0.505
Indirect 1 ( $a_1 * a_2$ )	0.055	0.036	0.126	-0.010, 0.133
Indirect 2 ( $b_1 * b_2$ )	-0.142	0.039	<0.001	-0.225, -0.076
Indirect 3 ( $b_1 * d_{21} * b_2$ )	-0.009	0.010	0.363	-0.031, 0.009
Total indirects	-0.096	0.052	0.067	-0.205, 0.004
Contrast1 (Ind1 - Ind2)	0.197	0.053	<0.001	0.101, 0.308
Contrast2 (Ind1 - Ind3)	0.064	0.039	0.103	-0.009, 0.153
Contrast3 (Ind2 - Ind3)	-0.133	0.041	0.001	-0.225, -0.06

---

Note. GRMS = gendered racial microaggressions. The significance of the indirect effects was calculated with bootstrapped, bias-corrected, confidence intervals (.95).

---

Working through the data, we should be able to find these items:

- The model accounts for 37% of the variance in predicting mental health outcomes.

- The total effect of GRMS (X) on mental health (Y) is  $-0.631, (p < .001)$ ; it is negative and statistically significant.
- The direct effect of GRMS (X) on mental health (Y) ( $-0.535, p < 0.001$ ) is still negative. Although someone lower in magnitude, it is still statistically significant. While inconsistent with the Baron and Kenny [1986] logic of mediation, Hayes [Hayes, 2022b] argues that a statistically significant indirect effect can stand on its own.
- Indirect effect #1 ( $a_1 \times b_1$  or GRMS through engagement coping to mental health) is  $B = 0.055, p = 0.126$ . As in the parallel mediation,  $p$  is  $> .05$  and the 95% CIs pass through zero ( $-0.010, 0.133$ ). Examining the individual paths, there is a statistically significant relationship from GRMS to engagement, but not from engagement to mental health.
- Indirect effect #2 ( $a_2 \times b_2$ , or GRMS through disengagement coping to mental health, is  $B = -0.142, p < 0.001, 95CI(-0.225, -0.076)$ ). Each of the paths is statistically significant from zero and so is the indirect effect.
- Indirect effect #3 ( $a_2 \times d_{21} \times b_2$ ; GRMS through engagement coping through disengagement coping to mental health) is  $-0.009, p = 0.363, 95C(-0.031, 0.009)$ . This indirect effect involves  $a_1$  (GRMS to engagement) and  $b_2$  which are significant. However, the path from engagement coping to disengagement coping is not significant.
- Total indirect:  $B = -0.096, p = 0.067$  is the sum of all specific indirect effects and is not statistically significant. The positive and negative indirects likely cancel each other out.
- With **contrasts** we ask: Are the indirect effects statistically significantly different from each other?
  - Contrast 1 (indirect 1 v 2):  $B = 0.197, p < 0.001$ , yes
  - Contrast 2 (indirect 1 v 3):  $B = 0.064, p = 0.103$ , no
  - Contrast 3 (indirect 2 v 3):  $B = -0.133, p = 0.001p$ , yes
  - This formal test of contrasts is an important one. It is not ok to infer that effects are statistically significantly different than each other on the basis of their estimates or  $p$  values. The formal test allows us to claim (with justification) that there are statistically significant differences between indirect effects 1 and 2; and 2 and 3.

### 6.5.3 APA Style Writeup

#### Method

**Data Analysis** Serial multiple mediation is appropriate when testing the influence of an independent variable (X) on the dependent variable (Y) directly, as well as indirectly through two or more mediators (M) and there is reason to hypothesize that variables that are causally prior in the model affect all variables later in the causal sequence [Hayes, 2022b]. We utilized serial multiple mediation analysis to test the influence of gendered racial microaggressions (X, GRMS) on mental health (Y, MntlHlth) directly as well as indirectly through the mediators engagement coping (M1, Engmt) and disengagement coping (M2, DisEngmt). Moreover, we hypothesized a causal linkage between the engagement coping mediator to the disengagement coping mediator such that a third specific indirect effect began with GRMS (X) through engagement coping (M1) through disengagement coping (M2) to mental health (Y). Using the *lavaan* (v. 0.6-16) package in R we followed the procedures outlined in Hayes [2022b] by analyzing the strength and significance of four sets of effects: specific indirect, the total indirect, the direct, and total. Bootstrap analysis, a nonparametric sampling procedure, was used to test the significance of the indirect effects.

*Hayes would likely recommend that we say this with fewer acronyms and more words/story.*

**Results Preliminary Analyses** Descriptive statistics were computed, and all variables were assessed univariate normality. *You would give your results regarding skew, kurtosis, Shapiro Wilks', here. If relevant, you could also describe multivariate normality.* A summary of descriptive statistics and a correlation matrix for the study is provided in Table 1. These bivariate relations provide evidence to support the test of mediation analysis.

**Serial Multiple Mediation Analysis** A model of serial multiple mediation was analyzed examining the degree to which engagement and disengagement coping mediated the relationship between gendered racial microaggressions and mental health outcomes. Hayes [2022b] recommended this strategy over simple mediation models because it allows for all mediators to be examined, simultaneously and allows the testing of the seriated effect of prior mediators onto subsequent ones. Using the *lavaan* (v. 0.6-16) package in R, coefficients for specific indirect, total indirect, direct, and total were computed. Path coefficients refer to regression weights, or slopes, of the expected changes in the dependent variable given a unit change in the independent variables.

Results (depicted in Figure # and presented in Table #) suggest that 37% of the variance in behavioral intentions is accounted for by the three variables in the model. Two of the specific indirect effects were significant and were statistically significantly different from each other. Specifically, the effect of gendered racial microaggressions through disengagement coping to mental health ( $B = -0.142, SE = 0.039, p < .001, 95CI[-0.076, -0.142]$ ) was stronger than the indirect effect from gendered racial microaggressions through engagement coping through disengagement coping to mental health ( $B = 0.055, SE = 0.036, p = 0.126, 95CI[0.133, 0.055]$ ). Interpreting the results suggests that, mental health outcomes are negatively impacted by gendered racial microaggressions direct and indirectly through disengagement coping. It is this latter path that has the greatest impact.

*Note:* In a manner consistent with the Lewis et al. [2017] article, the APA Results section can be fairly short. This is especially true when a well-organized table presents the results. In fact, I could have left all the numbers out of this except for the  $R^2$  (because it was not reported in the table).

## 6.6 STAY TUNED

A section on power analysis is planned and coming soon! My apologies that it's not quite Ready.

## 6.7 Troubleshooting and FAQs

An indirect effect that was (seemingly) significant in a simple (single) mediation disappears when additional mediators are added.

- Correlated mediators (e.g., multicollinearity) is a likely possibility.
- Which is correct? Maybe both...

A total effect was not significant, but there is one or more statistically significant specific indirect effect

- Recall that a total effect equals the sum of direct and indirect effects. If one specific indirect effect is positive and another is negative, this could account for the NS total effect.

- If the direct effect is NS, but the indirect effects are significant, this might render the total effect NS.
- The indirect effects might operate differently in subpopulations (males, females).

Your editor/peer reviewer/dissertation chair-or-committee member may insist that you do this the Baron & Kenny way (aka “the causal steps approach”).

- Hayes [Hayes, 2022a] provides compelling arguments for how to justify your (I believe correct) decision to just use the PROCESS (aka, bootstrapped, bias corrected, CIs )approach.
- My favorite line in his text reads, ” (the Baron and Kenny way)...is still being taught and recommended by researchers who don’t follow the methodology literature.”

How can I extend a mediation (only) model to include multiple Xs, Ys, or COVs?

- There is fabulous, fabulous narration and syntax for doing all of this in Hayes text. Of course his mechanics are in PROCESS, but *lavaan* is easy to use by just “drawing more paths” via the syntax. We’ll get more practice as we go along.

What about effect sizes? Shouldn’t we be including/reporting them?

- Yes! The closest thing we have reported to an effect size is  $R^2$ , which assess proportion of variance accounted for in the M and Y variables.
- In PROCESS and path analysis this is still emerging. Hayes chapter 4 presents a handful of options for effect sizes beyond  $R^2$ .

## 6.8 Practice Problems

The three problems described below are designed to be grow in this series of chapters that begins with simple mediation and progresses through complex mediation, moderated moderation, and conditional process analysis. The goal of this assignment is to conduct a complex (e.g., parallel or serial) mediation.

I recommend that you select a dataset that includes at least four variables. If you are new to this topic, you may wish to select variables that are all continuously scaled. The IV and moderator (in subsequent chapters) *could* be categorical (if they are dichotomous, please use 0/1 coding; if they have more than one category it is best if they are ordered). You will likely encounter challenges that were not covered in this chapter. Search for and try out solutions, knowing that there are multiple paths through the analysis.

The suggested practice problem for this chapter is to conduct a parallel or serial mediation (or both).

### 6.8.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed

If conducting a parallel or serial mediation feels a bit overwhelming, simply change the random seed in the data simulation, then rework one of the chapter problems (i.e., parallel or serial mediation). This should provide minor changes to the data (maybe in the second or third decimal point), but the results will likely be very similar.

### 6.8.2 Problem #2: Rework the research vignette, but swap one or more variables

Conduct the complex mediation (parallel or serial) using the simulated data provided in this chapter, but swap out one or more of the variables. This could mean changing roles for the variables that were the focus of the chapter, or substituting one or more variables for those in the simulated data but not modeled in the chapter.

### 6.8.3 Problem #3: Use other data that is available to you

To conduct the parallel or serial mediation, use data for which you have permission and access. This could be IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; or data from other chapters (or the “homeworked example”) in this OER.

### 6.8.4 Grading Rubric

Assignment Component		
1. Assign each variable to the X, Y, M1, and M2 roles	5	_____
4. Use tidySEM to create a figure that represents your results	5	_____
5. Create a table that includes a summary of the effects (indirect, direct, total, total indirect) as well as contrasts	5	_____
6. Represent your work in an APA-style write-up	5	_____
7. Explanation to grader	5	_____
8. Be able to hand-calculate the indirect, direct, and total effects from the a, b, & c' paths	5	_____
<b>Totals</b>	<b>40</b>	_____

## 6.9 Homeworked Example

### Screencast Link

For more information about the data used in this homeworked example, please refer to the description and codebook located at the end of the [introductory lesson](#) in [ReCentering Psych Stats](#). An .rds file which holds the data is located in the [Worked Examples](#) folder at the GitHub site the hosts the OER. The file name is *ReC.rds*.

The suggested practice problem for this chapter is to conduct a complex (i.e., parallel or serial) mediation.

### Assign each variable to the X, Y, M1, and M2 roles

X = Centering: explicit recentering (0 = precentered; 1 = recentered) M1 = TradPed: traditional pedagogy (continuously scaled with higher scores being more favorable) M2 = SRPed: socially responsive pedagogy (continuously scaled with higher scores being more favorable) Y = Valued: valued by me (continuously scaled with higher scores being more favorable)

In this *parallel mediation*, I am hypothesizing that the perceived course value to the students is predicted by intentional recentering through their assessments of traditional and socially responsive pedagogy.

It helps me to make a quick sketch:

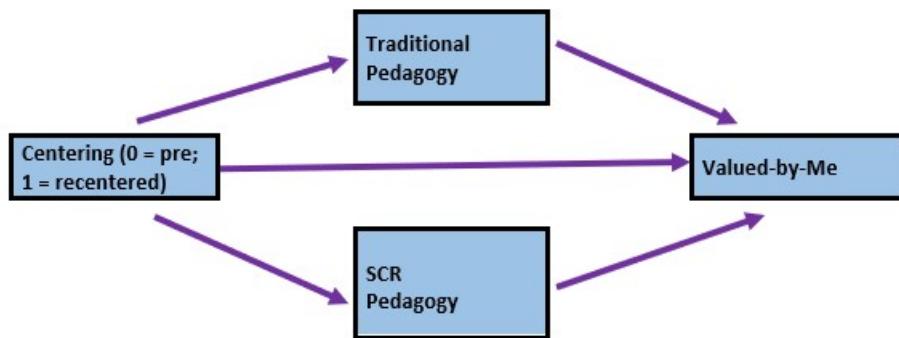


Figure 6.4: An image of the parallel mediation model for the homeworked example.

### Import the data and format the variables in the model

```
raw <- readRDS("ReC.rds")
```

The approach we are taking to complex mediation does not allow dependency in the data. Therefore, we will include only those who took the multivariate class (i.e., excluding responses for the ANOVA and psychometrics courses).

```
raw <- (dplyr::filter(raw, Course == "Multivariate"))
```

I need to score the TradPed, SRPed, and Valued variables

```
TradPed_vars <- c("ClearResponsibilities", "EffectiveAnswers", "Feedback",
                  "ClearOrganization", "ClearPresentation")
raw$TradPed <- sjstats::mean_n(raw[, ..TradPed_vars], 0.75)
```

```
Valued_vars <- c("Val0bjectives", "IncrUnderstanding", "IncrInterest")
raw$Valued <- sjstats::mean_n(raw[, ..Valued_vars], 0.75)

SRPed_vars <- c("InclusvClassrm", "EquitableEval", "MultPerspectives",
  "DEIintegration")
raw$SRPed <- sjstats::mean_n(raw[, ..SRPed_vars], 0.75)
```

I will create a babydf.

```
babydf <- dplyr::select(raw, Centering, TradPed, Valued, SRPed)
```

Let's check the structure of the variables:

```
str(babydf)
```

```
## Classes 'data.table' and 'data.frame': 84 obs. of 4 variables:
## $ Centering: Factor w/ 2 levels "Pre","Re": 2 2 2 2 2 2 2 2 2 ...
## $ TradPed : num 3.8 5 4.8 4 4.2 3 5 4.6 4 4.8 ...
## $ Valued : num 4.33 5 4.67 3.33 4 3.67 5 4 4.67 4.67 ...
## $ SRPed : num 4.5 5 5 5 4.75 4.5 5 4.5 5 5 ...
## - attr(*, ".internal.selfref")=<externalptr>
```

At this point, these my only inclusion/exclusion criteria. I can determine how many students (who consented) completed any portion of the survey.

## Specify and run the lavaan model

```
ReCpMed <- "
  Valued ~ b1*TradPed + b2*SRPed + c_p*Centering
  TradPed ~ a1*Centering
  SRPed ~ a2*Centering

  indirect1 := a1 * b1
  indirect2 := a2 * b2
  contrast := indirect1 - indirect2
  total_indirects := indirect1 + indirect2
  total_c     := c_p + (indirect1) + (indirect2)
  direct := c_p
  "

ReCpMedfit <- lavaan::sem(ReCpMed, data = babydf, se = "bootstrap", missing = "fiml")
ReCpMedsummary <- lavaan::summary(ReCpMedfit, standardized = T, rsq = T,
  fit = TRUE, ci = TRUE)
ReC_pMedParamEsts <- lavaan::parameterEstimates(ReCpMedfit, boot.ci.type = "bca.simple",
  standardized = TRUE)
ReCpMedsummary
```

```
## lavaan 0.6.16 ended normally after 23 iterations
##
##   Estimator                      ML
## Optimization method            NLMINB
## Number of model parameters    11
##
##   Number of observations        84
## Number of missing patterns     3
##
## Model Test User Model:
##
##   Test statistic                 54.059
##   Degrees of freedom              1
##   P-value (Chi-square)           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                 145.642
##   Degrees of freedom                6
##   P-value                          0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)      0.620
##   Tucker-Lewis Index (TLI)          -1.280
##
##   Robust Comparative Fit Index (CFI) 0.613
##   Robust Tucker-Lewis Index (TLI)    -1.323
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)    -202.536
##   Loglikelihood unrestricted model (H1) -175.506
##
##   Akaike (AIC)                     427.071
##   Bayesian (BIC)                   453.810
##   Sample-size adjusted Bayesian (SABIC) 419.110
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                           0.795
##   90 Percent confidence interval - lower 0.623
##   90 Percent confidence interval - upper 0.982
##   P-value H_0: RMSEA <= 0.050       0.000
##   P-value H_0: RMSEA >= 0.080       1.000
##
##   Robust RMSEA                     0.815
##   90 Percent confidence interval - lower 0.641
```

```

## 90 Percent confidence interval - upper          1.004
## P-value H_0: Robust RMSEA <= 0.050          0.000
## P-value H_0: Robust RMSEA >= 0.080          1.000
##
## Standardized Root Mean Square Residual:
##
## SRMR                               0.217
##
## Parameter Estimates:
##
## Standard errors                      Bootstrap
## Number of requested bootstrap draws   1000
## Number of successful bootstrap draws  1000
##
## Regressions:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## Valued ~
##   TradPed (b1)    0.686   0.133   5.168   0.000   0.470   0.976
##   SRPed   (b2)    0.119   0.138   0.867   0.386  -0.176   0.380
##   Centerng (c_p)  0.015   0.101   0.145   0.885  -0.174   0.214
## TradPed ~
##   Centerng (a1)  0.312   0.146   2.135   0.033   0.014   0.588
## SRPed ~
##   Centerng (a2)  0.353   0.120   2.931   0.003   0.107   0.577
## Std.lv Std.all
##
##   0.686   0.747
##   0.119   0.104
##   0.015   0.011
##
##   0.312   0.210
##
##   0.353   0.296
##
## Intercepts:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .Valued      0.710   0.474   1.498   0.134  -0.203   1.671
## .TradPed     3.870   0.244  15.865   0.000   3.409   4.358
## .SRPed       4.029   0.196  20.507   0.000   3.643   4.409
## Std.lv Std.all
##
##   0.710   1.077
##   3.870   5.396
##   4.029   7.013
##
## Variances:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .Valued      0.181   0.028   6.386   0.000   0.117   0.227
## .TradPed     0.492   0.128   3.826   0.000   0.253   0.767

```

```

##      .SRPed          0.301    0.059    5.071    0.000    0.191    0.419
##      Std.lv  Std.all
##      0.181    0.418
##      0.492    0.956
##      0.301    0.912
##
## R-Square:
##           Estimate
##      Valued      0.582
##      TradPed     0.044
##      SRPed       0.088
##
## Defined Parameters:
##           Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##      indirect1    0.214   0.108   1.986   0.047    0.008    0.439
##      indirect2    0.042   0.053   0.794   0.427   -0.073    0.152
##      contrast     0.172   0.125   1.373   0.170   -0.035    0.454
##      total_indircts  0.256   0.115   2.231   0.026    0.028    0.473
##      total_c      0.271   0.145   1.875   0.061   -0.012    0.551
##      direct       0.015   0.101   0.145   0.885   -0.174    0.214
##      Std.lv  Std.all
##      0.214    0.157
##      0.042    0.031
##      0.172    0.126
##      0.256    0.188
##      0.271    0.199
##      0.015    0.011

```

## ReC\_pMedParamEsts

	lhs	op	rhs	label	est	se
## 1	Valued	~	TradPed	b1	0.686	0.133
## 2	Valued	~	SRPed	b2	0.119	0.138
## 3	Valued	~	Centering	c_p	0.015	0.101
## 4	TradPed	~	Centering	a1	0.312	0.146
## 5	SRPed	~	Centering	a2	0.353	0.120
## 6	Valued	~~	Valued		0.181	0.028
## 7	TradPed	~~	TradPed		0.492	0.128
## 8	SRPed	~~	SRPed		0.301	0.059
## 9	Centering	~~	Centering		0.233	0.000
## 10	Valued	~1			0.710	0.474
## 11	TradPed	~1			3.870	0.244
## 12	SRPed	~1			4.029	0.196
## 13	Centering	~1			1.369	0.000
## 14	indirect1	:=	a1*b1	indirect1	0.214	0.108
## 15	indirect2	:=	a2*b2	indirect2	0.042	0.053
## 16	contrast	:=	indirect1-indirect2	contrast	0.172	0.125

```

## 17 total_indirects := indirect1+indirect2 total_indirects 0.256 0.115
## 18      total_c := c_p+(indirect1)+(indirect2)      total_c 0.271 0.145
## 19      direct := c_p      direct 0.015 0.101
##          z pvalue ci.lower ci.upper std.lv std.all std.nox
## 1   5.168 0.000    0.421    0.920  0.686   0.747   0.747
## 2   0.867 0.386   -0.135    0.416  0.119   0.104   0.104
## 3   0.145 0.885   -0.176    0.209  0.015   0.011   0.022
## 4   2.135 0.033    0.014    0.591  0.312   0.210   0.435
## 5   2.931 0.003    0.098    0.568  0.353   0.296   0.614
## 6   6.386 0.000    0.138    0.253  0.181   0.418   0.418
## 7   3.826 0.000    0.282    0.835  0.492   0.956   0.956
## 8   5.071 0.000    0.212    0.458  0.301   0.912   0.912
## 9     NA    NA    0.233    0.233  0.233   1.000   0.233
## 10  1.498 0.134   -0.174    1.732  0.710   1.077   1.077
## 11 15.865 0.000    3.364    4.326  3.870   5.396   5.396
## 12 20.507 0.000    3.625    4.402  4.029   7.013   7.013
## 13    NA    NA    1.369    1.369  1.369   2.837   1.369
## 14  1.986 0.047    0.008    0.439  0.214   0.157   0.325
## 15  0.794 0.427   -0.028    0.220  0.042   0.031   0.064
## 16  1.373 0.170   -0.036    0.453  0.172   0.126   0.261
## 17  2.231 0.026    0.038    0.482  0.256   0.188   0.389
## 18  1.875 0.061   -0.017    0.539  0.271   0.199   0.411
## 19  0.145 0.885   -0.176    0.209  0.015   0.011   0.022

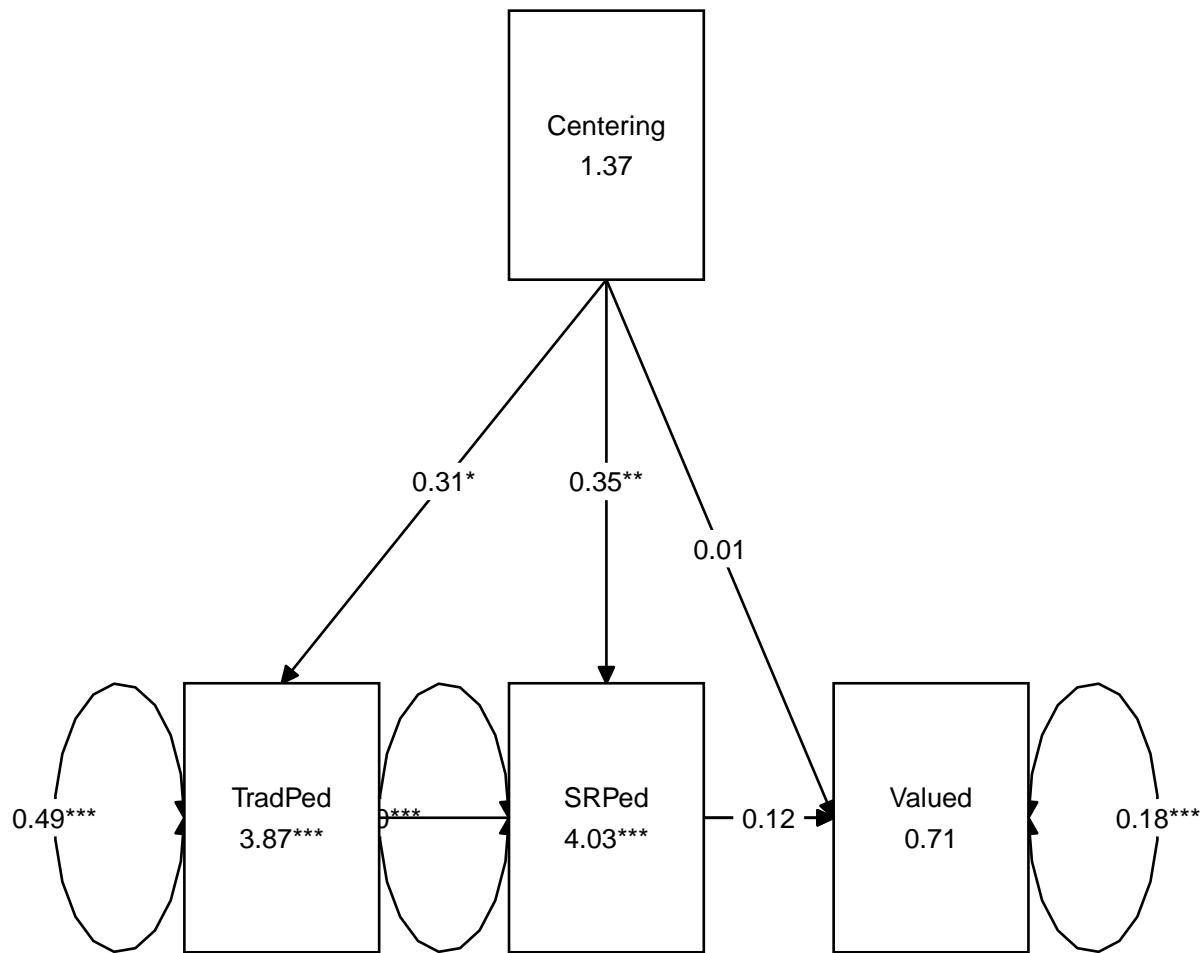
```

Use tidySEM to create a figure that represents your results

```

# only worked when I used the library to turn on all these pkgs
library(lavaan)
library(dplyr)
library(ggplot2)
library(tidySEM)
tidySEM::graph_sem(model = ReCpMedfit)

```



```
tidySEM::get_layout(ReCpMedfit)
```

```
##      [,1]      [,2]      [,3]
## [1,] NA "Centering" NA
## [2,] "TradPed" "SRPed" "Valued"
## attr(),"class")
## [1] "layout_matrix" "matrix"      "array"
```

To create the figure I showed at the beginning of the chapter, we will want three rows and three columns.

```
ReCpMed_map <- tidySEM::get_layout("", "TradPed", "", "Centering", "", 
  "Valued", "", "SRPed", "", rows = 3)
ReCpMed_map
```

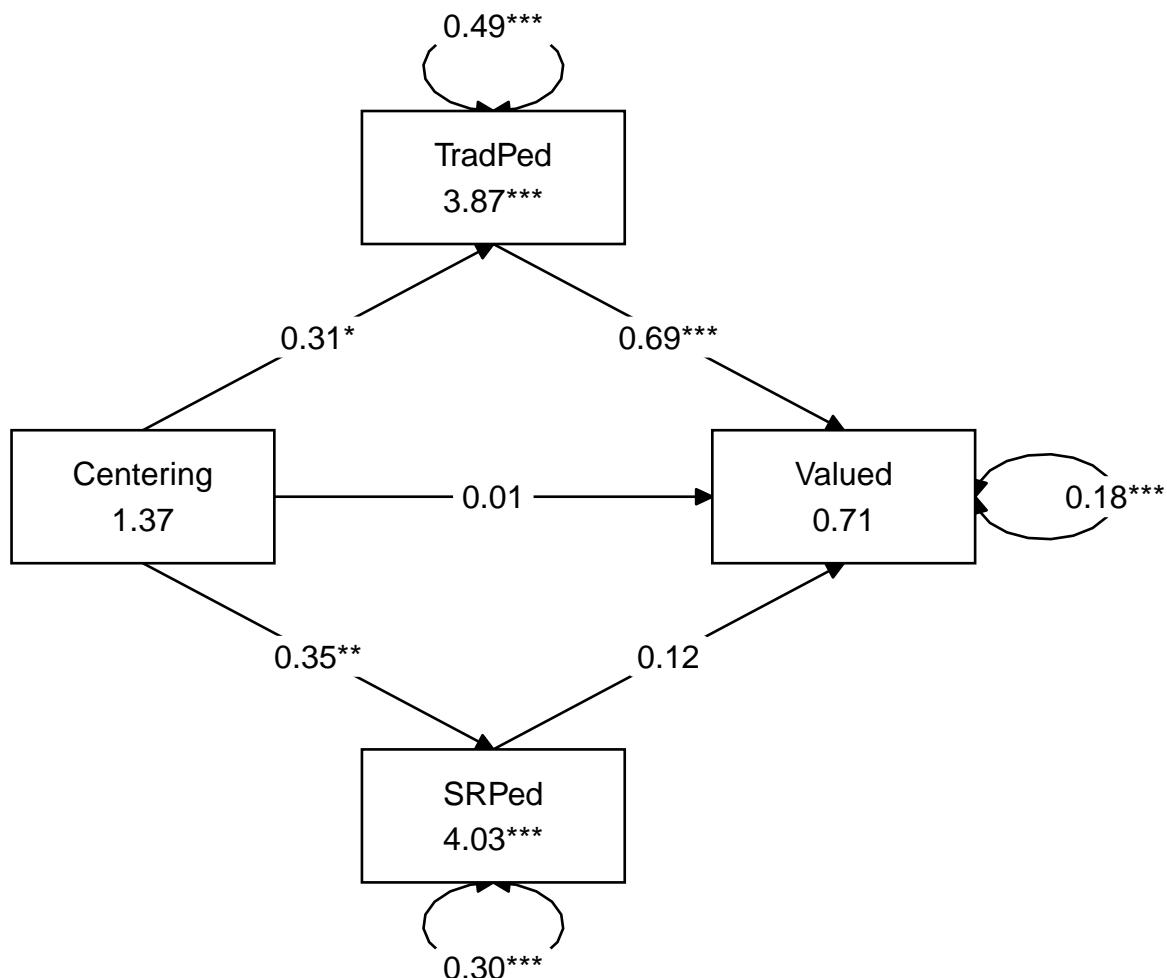
```
##      [,1]      [,2]      [,3]
```

```

## [1] ""           "TradPed"   ""
## [2] "Centering" ""          "Valued"
## [3] ""           "SRPed"    ""
## attr(),"class")
## [1] "layout_matrix" "matrix"      "array"

tidySEM::graph_sem(ReCpMedfit, layout = ReCpMed_map, rect_width = 1.5,
rect_height = 1.25, spacing_x = 2, spacing_y = 3, text_size = 4.5)

```



Create a table that includes a summary of the effects (indirect, direct, total, total indirect) as well as contrasts

I will write my results to a .csv file.

```
write.csv(ReC_pMedParamEsts, file = "ReC_pMedParamEsts.csv")
```

**Table 1**

Model Coefficients Assessing Students' Appraisal of Traditional and Socially Responsive Pedagogy in a Model of Parallel Mediation Predicting Perceived Course Value from Explicit Recentering

Predictor	<i>B</i>	<i>SE<sub>B</sub></i>	<i>p</i>	<i>R</i> <sup>2</sup>
Traditional Pedagogy (M1)				
Constant	3.870	0.234	<0.001	
Centering ( <i>a</i> <sub>1</sub> )	0.312	0.141	0.027	
Socially Responsive Pedagogy (M2)				
Constant	4.029	0.193	<0.001	.09
Centering ( <i>a</i> <sub>2</sub> )	0.353	0.116	0.002	
Perceived Course Value (DV)				
Constant	0.710	0.477	0.136	
Traditional Pedagogy ( <i>b</i> <sub>1</sub> )	0.686	0.133	<0.001	
Socially Rx Pedagogy ( <i>b</i> <sub>2</sub> )	0.119	0.141	0.397	
Centering ( <i>c'</i> )	0.015	0.102	0.885	
Effects				
Total effect	0.271	0.143	0.059	-0.024, 0.550
Indirect 1 ( <i>a</i> <sub>1</sub> * <i>b</i> <sub>1</sub> )	0.214	0.103	0.037	0.035, 0.440
Indirect 2 ( <i>a</i> <sub>2</sub> * <i>b</i> <sub>2</sub> )	0.042	0.053	0.429	-0.040, 0.184
Total indirects	0.256	0.111	0.021	0.060, 0.489
Contrast1 (Ind1 - Ind2)	0.172	0.120	0.152	-0.041, 0.423

*Note.* The significance of the indirect effects was calculated with bootstrapped, bias-corrected, confidence intervals (.95).

### Represent your work in an APA-style write-up

A model of parallel mediation analyzed the degree to which students' perceptions of traditional and socially responsive pedagogy mediated the relationship between explicit recentering of the course and course value. Hayes [2022b] recommended this strategy over simple mediation models because it allows for all mediators to be examined, simultaneously. The resultant direct and indirect values for each path account for other mediation paths. Using the *lavaan* (v. 0.6-16) package in R, coefficients for specific indirect, total indirect, direct, and total were computed. Path coefficients refer to regression weights, or slopes, of the expected changes in the dependent variable given a unit change in the independent variables.

Results (depicted in Figure 1 and presented in Table 1) suggest that 58% of the variance in perceptions of course value is accounted for by the model. The indirect effect predicting course value from explicit recentering through traditional pedagogy was statistically significant ( $B = 0.214, SE = 0.103, p = 0.037, 95CI[0.035, 0.440]$ ). Examining the individual paths we see that  $a_1$  was positive and statistically significant (recentering is associated with higher evaluations of traditional pedagogy). The  $b_1$  path was similarly statistically significant (traditional pedagogy was associated with course valuation). The indirect effect predicting course value from recentering through socially responsive pedagogy was not statistically significant  $B = 0.042, SE = 0.053, p = 0.429, 95CE[-0.040, 0.184]$ ). While explicit recentering had a statistically significant effect on ratings of socially responsive pedagogy (i.e., the  $a_2$  path), socially responsive pedagogy did not have a statistically significant effect on perceptions of course value (i.e., the  $b_2$  path). The drop in magnitude and near-significance from the total effect ( $B = 0.271, p = 0.059$ ) to the direct effect ( $B = 0.015, p = 0.885$ ) supports the presence of mediation. A pairwise comparison of the specific indirect effects indicated that the strength of the effects were not statistically significantly different from each other. In summary, the effects of explicit recentering on perceived value to the student appears to be mediated through students evaluation of traditional pedagogy.

### Explanation to grader

**Be able to hand-calculate the indirect, direct, and total effects from the a, b, & c' paths**

- Indirect =  $a^*b$
- Direct = Total minus indirect
- Total =  $(a^*b) + c'$

### A homework idea

Augment this model to a serial mediation – adding a path from traditional pedagogy to socially responsive pedagogy.

# **MODERATION**



# Chapter 7

## Simple Moderation in OLS and MLE

### [Screencasted Lecture Link](#)

The focus of this lecture is an overview of simple moderation. Sounds simple? Wait, there's more! The focus of this lecture is the transition:

- from null hypothesis significance testing (NHST) to modeling
- from *ordinary least squares* (OLS) to *maximum likelihood estimation* (MLE)

In making the transition we will work a moderation/interaction problem from Hayes' text with both `lm()` and `lavaan/sem()` functions.

### 7.1 Navigating this Lesson

There is about 1 hour and 10 minutes of lecture. If you work through the materials with me it would be plan for an additional hour

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

#### 7.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Distinguish between NHST and model building approaches
- Name the primary characteristics that distinguish ordinary least squares from maximum likelihood approaches to regression.
- Interpret “the usual” things we find in regression: B/beta weights, R,  $R^2$ .
- Define and interpret simple slopes and probing an interaction, this includes
  - pick-a-point and Johnson-Neyman approaches

- interpreting interaction plots/figures
- Recognize the path specification in *lavaan*. That is, you should be able to figure out a diagram from the *lavaan* code. In reverse, you should be able to write (or identify) the proper code in *lavaan*.

### 7.1.2 Planning for Practice

Although I provide more complete descriptions at the end of the chapter follow these suggestions, providing an overview of them here may help you plan for what you might want to do as you work through the chapter. As is typical for this OER, the suggestions for homework are graded in complexity. I recommend you select an option that builds on your confidence but provides a bit of stretch. I also suggest you utilize a dataset that has at least four variables that are suitable for growing into a complex moderation (additive or moderated) or moderated mediation. This will be easiest if the variables are continuous in nature. In these chapters, I do not describe how to use categorical variables in dependent (e.g., consequent or endogenous) roles. However, dichotomous and ordered factors are suitable as independent variables and covariates.

- Rework the problem in the chapter by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- There are a number of variables in the dataset. Swap out one or more variables in the simple moderation and compare your solution to the one in the chapter (and/or one you mimicked in the journal article).
- Conduct a simple moderation with data to which you have access. This could include data you simulate on your own or from a published article.

### 7.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

Regarding ordinary least squares (OLS) versus maximum likelihood estimation (MLE), these articles are extremely helpful:

- Cohen, J. (2003). Maximum likelihood estimation. Section 13.2.9 (pp. 498-499). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum Associates.
- Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), 90–100. [https://doi.org/10.1016/S0022-2496\(02\)00028-7](https://doi.org/10.1016/S0022-2496(02)00028-7) (skim for big ideas)
- Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, 65(1), 1–12. <https://doi.org/10.1037/a0018326>

Regarding the topic of moderation, I drew heavily from these resources.

- Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach*. Guilford Publications. <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=5109647>
  - Chapter 7: Fundamentals of Moderation Analysis: This chapter focuses on the basics of moderation analysis. Our goal is to transfer and apply the knowledge to models we run in lavaan. An excellent review of centering, visualizations, and probing moderations.
  - Chapter 8: Extending the Fundamental Principles of Moderation Analysis (pp. 267-301): Hayes addresses common regression concerns such as (a) hierarchical vs. simultaneous entry and (b) comparison of moderated regression with 2x2 factorial ANOVA.
  - Chapter 9: Some Myths and Additional Extensions of Moderation Analysis (pp. 303-347). Hayes identifies “truths and myths” about mean centering and standardization. For sure these are important topics and his take on them is clear and compelling.
  - Appendix A: An essential tool for PROCESS users because, even when we are in the R environment, this is the “idea book.” That is, the place where all the path models are presented in figures.

The research vignette for this chapter:

- Kim, P. Y., Kendall, D. L., & Cheon, H.-S. (2017). Racial microaggressions, cultural mistrust, and mental health outcomes among Asian American college students. *American Journal of Orthopsychiatry*, 87(6), 663–670. <https://doi-org.ezproxy.spu.edu/10.1037/ort0000203>

#### 7.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
if (!require(apaTables)) {
  install.packages("apaTables")
}
```

Loading required package: apaTables

```
if (!require(lavaan)) {
  install.packages("lavaan")
}
```

Loading required package: lavaan

```
This is lavaan 0.6-16
lavaan is FREE software! Please report any bugs.
```

```
if (!require(semPlot)) {
  install.packages("semPlot")
}
```

Loading required package: semPlot

```
if (!require(tidyverse)) {
  install.packages("tidyverse")
}
```

Loading required package: tidyverse

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr    1.3.0
v purrr    1.0.1
-- Conflicts -----
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
if (!require(psych)) {
  install.packages("psych")
}
```

Loading required package: psych

Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

%+%, alpha

The following object is masked from 'package:lavaan':

cor2cov

```
if (!require(jtools)) {
  install.packages("jtools")
}
```

Loading required package: jtools

## 7.2 On *Modeling*: Introductory Comments on the simultaneously invisible and paradigm-shifting transition we are making

### 7.2.1 NHST versus modeling

At least a decade old now, Rogers' [2010] article in the *American Psychologist* is one of my favorites. In it, he explores the notion of *statistical modeling*. He begins with criticisms of null hypothesis statistical testing by describing how it has become a awkward and incongruent blend of Fisherian (i.e., R.A. Fisher) and Neyman-Pearson (i.e., Jerzy Neyman and E. S. Pearson) approaches.

**Table 1**

Contributions of the Fisherian and Neyman-Pearson Approaches to NHST [Rodgers, 2010]	
Fisher	Neyman-Pearson
Developed NHST to answer scientific questions and evaluate theory. Took an incremental approach to hypothesis testing that involved replication and (potentially) self-correcting; as such viewed <i>replication</i> as a critical element. Never used the terms, “alternative hypothesis” or “alpha level.” Rather, Fisher used the distribution of the null model to examine “whether the data look weird or not.” Gave us the null hypothesis and $p$ value.	Sought to draw conclusions in applied settings such as quality control. Placed emphasis on the importance of each individual decision. Designed their approach to detect an “alternative hypothesis.” Gave us the alternative hypothesis, alpha level, and power.

Over time, these overlapping, but inconsistent, approaches became intertwined. Many students of statistics do not recognize the incompatibilities. Undoubtedly, it makes statistics more difficult to learn (and teach). Below are some of the challenges that Rodgers [2010] outlined.

- Rejecting the null does not provide logical or strong support for the alternative
- Failing to reject the null does not provide logical or strong support for the null.
- NHST is backwards because it evaluates the probability of the data given the hypothesis, rather than the probability of the hypothesis given the data.
- All point-estimate null hypotheses can be rejected if the sample size is large enough.
- Statistical significance does not necessitate practical significance.

Consequently, we have ongoing discussion/debates about power, effect sizes, sample size, Type I and II errors, confidence intervals, fit statistics, and the relations between them.

### 7.2.2 Introducing: *The Model*

Understanding modeling in our *scientist-practitioner* context probably needs to start with understanding the *mathematical model*. Niemark and Este [1967] defined a mathematical model as a set of assumptions together with implications drawn from them by mathematical reasoning. Luce [Luce, 1995] suggested that mathematical equations capture model-specific features by highlighting some aspects while ignoring others. The use of mathematics helps us uncover the “structure.” For example, the *mean* is a mathematical model. *I always like to stop and think about that notion...about what the mean represents and what it doesn’t.* Pearl [2000] defined the model as an idealized representation of reality that highlights some aspects and ignores others by suggesting that a model:

- matches the reality it describes in some important ways.
- is simpler than that reality.

As we transition from the NHST approach to statistical modeling there is [Rodgers, 2010]:

- decreased emphasis on
  - null hypothesis
  - $p$  values
- increased emphasis on
  - model residuals
  - degrees of freedom
  - additional indices of *fit*

Further, statistical models [Rodgers, 2010]:

- are more readily falsifiable
- require greater theoretical precision
- include assumptions that are more readily evaluated
- offer more practical application

Circling back around to Fisher and Neyman-Pearson, Rogers [2010] contended that Fisher’s work provided a framework for modeling because of the model process of specification, estimation, and goodness of fit. As we move into more complex modeling, we will spend a great deal of time understanding parameters and their relationship to degrees of freedom. Fisher viewed degrees of freedom as *statistical currency* that could be used in exchange for the estimation of parameters.

If this topic is exciting to you, let me refer you to Cumming’s [Cumming, 2014] article, “The New Statistics: Why and How,” in the Journal, \*Psychological Science”

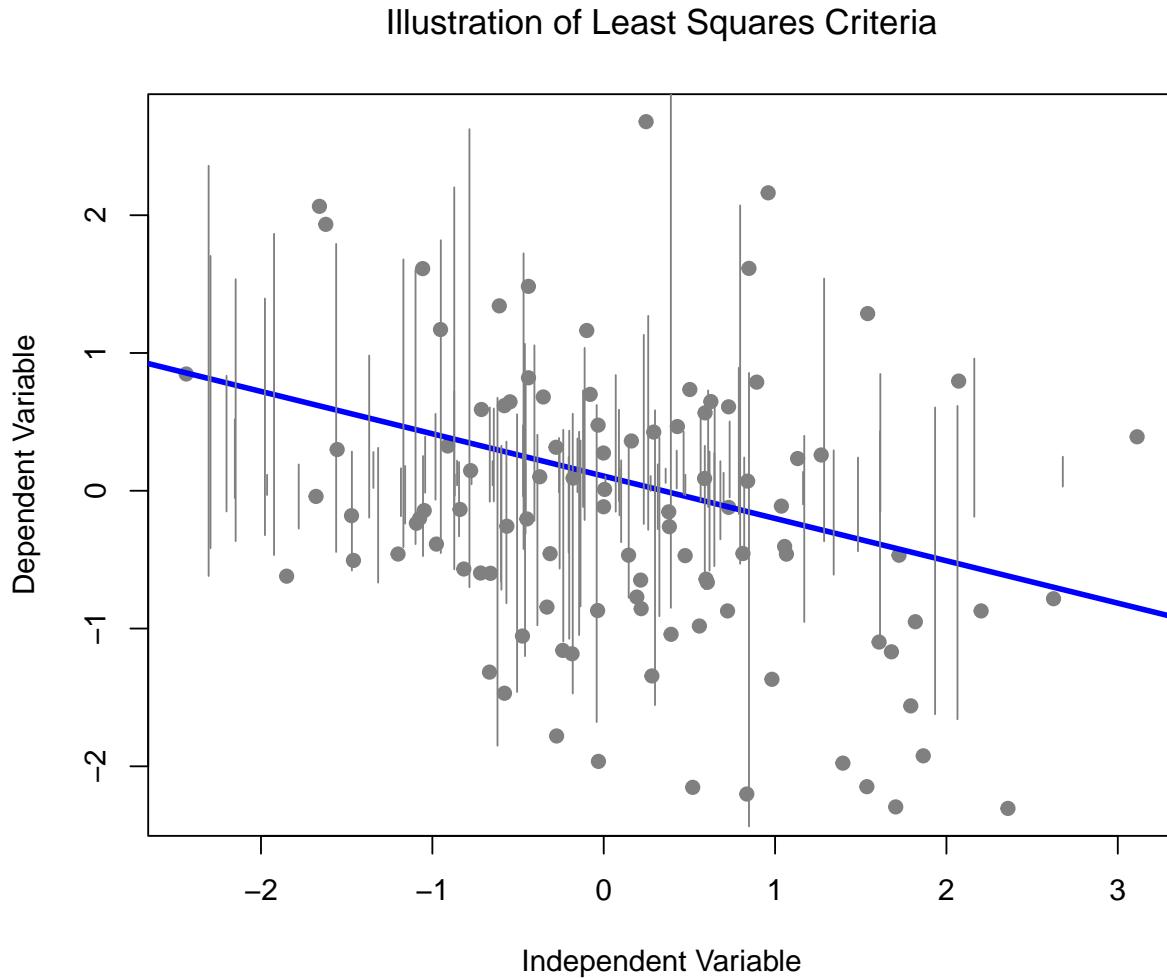
## 7.3 OLS to ML for Estimation

### 7.3.1 Ordinary least squares (OLS)

Known by a variety of names, the estimation algorithm typically used in regression models (linear, hierarchical, multiple, sequential) is *ordinary least squares* (OLS; also termed least squares criterion, general least squares, etc.). As we move into multivariate (and then psychometrics) we are going to transition our estimation method from OLS to MLE. Consequently, it is essential to understand some underlying differences [Cohen et al., 2003, Myung, 2003]

In OLS regression:

- The estimated values of regression coefficients are chosen so that the sum of squared errors is minimized (aka, the *least squares criteria*). Consequently,
  - the mean of errors is zero, and
  - the errors correlate *zero* with each predictor
- The solution to OLS regression is *analytic*
  - the equations from which the coefficients are created are *known normal equations*. Among other places, you can look them up in CCW&A [Cohen and Nagel, 1934] Appendix 1)



### 7.3.2 Maximum likelihood estimation (MLE): A brief orientation

Although I started this chapter with a critique of NHST, Fisher is credited [Myung, 2003] with the original development of the central principle of *maximum likelihood estimation* which is that the desired probability distribution is the one that makes the observed data *most likely*. As such, the *MLE estimate* is a resulting parameter vector that maximizes the likelihood function. Myung's [2003] tutorial provides an excellent review. My summary is derived from Dr. Myung article. A *likelihood* is a measure of how *typical* a person (or sample) is of that population.

- When there is one IV the MLE distribution behaves like a chi-square distribution (which also tests observed versus expected data).
- There is a point in the MLE curve that represents where the maximum likelihood exists that the data is likely given the model.
- When there are multiple IVs, this simple curve takes the shape of a  $k$  dimensional geometrical surface.

Extended to regression, we are interested in the *likelihoods* of individuals having particular scores on Y, given values on predictors  $x_1$  to  $x_k$  (and the specific values of regression coefficients chosen as the parameter estimates)

- MLE provides *maximum likelihood estimates* of the regression coefficients (and SEs) that is, estimates that make a sample as likely or typical as possible
- $L$  is a symbol for *maximum likelihood of a sample*
- The solutions are *iterative* (i.e., identified by trial-and-error; with each trial informed by the prior)
  - a statistical criteria is specified for the coefficients to be chosen
  - different values of coefficients are tried
  - these *iterations* continue until the regression coefficients cease to change by more than a small amount (i.e., the *convergence criteria*)
  - hopefully, a set of coefficients is found that makes the solution as close to the statistical criteria (i.e., maximum likelihood) as possible
- The *optimization algorithm* does not guarantee that a set of parameters will be found; convergence failures may be caused by
  - multicollinearity among predictors
  - a large number of predictors
  - the *local maxima problem*; the optimization algorithm returns sub-optimal parameter values [Myung, 2003]
- MLE is a *full information model*
  - calculates the estimates of model parameters all at once
- MLE is for large samples
- MLE assumptions include
  - independence of observations
  - multivariate normality of endogenous variables
  - independence of exogenous variables and disturbances
  - correct specification of the model (MLE is only appropriate for testing theoretically informed models)

### 7.3.3 OLS and MLE Comparison

In this table we can compare OLS and MLE in a side-by-side manner. **Table 2**

---

Comparing OLS and MLE [Cohen et al., 2003, Myung, 2003]

---

Criterion	Ordinary Least Squares (OLS)	Maximum Likelihood Estimation (MLE)
-----------	------------------------------	-------------------------------------

---

Parameter values chosen to...	minimize the distance between the predictions from regression line and the observations; considered to be those that are <i>most accurate</i>	be those that are <i>most likely</i> to have produced the data
Parameter values are obtained by	equations that are known and linear (you can find them in the “back of the book”)	a non-linear optimization algorithm
Preferred when...	sample size is small	sample size is large, for complex models, non-linear models, and when OLS and MLE results differ
In R...	the <i>lm()</i> function in base R	<i>lavaan</i> and other packages*; specifying the FIML option allows for missing data (without imputation)

---

### 7.3.4 Hayes and PROCESS (aka conditional process analysis)

In the early 2000s, the bias-corrected, bootstrapped, confidence interval (CI) was identified as a more powerful approach to assessing indirect effects than the classic Sobel test. Because programs did not produce them, no one was using them. Preacher, Edwards, Lambert, Hayes, and colleagues created Excel worksheets that would calculate these (they were so painful). Hayes turned this process into a *series* of macros to do a variety of things for SPSS and other programs. Because of his clear, instructional, text, PROCESS is popular. In 2021, Hayes released the PROCESS macro for R. It can be downloaded at the [ProcessMacro website](#). Documentation for it is newly emerging. Although PROCESS produces bias-corrected, bootstrapped confidence intervals, for models with indirect effects, PROCESS utilizes OLS as the estimator.

Although most regression models can be completed with the *lm()* function in base R, it can be instructive to run a handful of these familiar models with *lavaan* (or even PROCESS) as a precursor to more complicated models.

## 7.4 Introducing the *lavaan* package

In the regression classes (as well as in research designs that are cross-sectional, non-linear, and can be parsimoniously and adequately measured with OLS regression) we typically use the base R function, *lm()* (“linear model”) which relies on an OLS algorithm. You can learn about it with this simple code:

```
#?lm
```

Rosseel’s [2020] *lavaan* package was developed for SEM, but is readily adaptable to most multiple regression models. Which do we use and when?

- For relatively simple models that involve only predictors, covariates, and moderators, *lm()* is adequate.
- Models that involve mediation need to use *lavaan*

- SEM/CFA needs *lavaan*
- If your sample size is small, *but* you are planning a mediation, it gets tricky (try to increase your sample size) because MLE estimators rely on large sample sizes (how big? hard to say).

### 7.4.1 The FIML magic for which we have been waiting

There are different types of maximum likelihood. In this chapter we'll utilize *full information maximum likelihood* (FIML). FIML is one of the most practical missing data estimation approaches around and is especially used in SEM and CFA. When data are thought to be MAR (missing at random) or MCAR (missing completely at random), it has been shown to produce unbiased parameter estimates and standard errors.

The FIML approach works by estimating a likelihood function for each individual based on the variables that are present so that all available data are used. Model fit is calculated from (or informed by) the fit functions for all individual cases. Hence, “FIML” is *full information* maximum likelihood.

When I am able to use *lavaan*, my approach is to use Parent's AIA (available information analysis, -Parent [2013]) approach to scoring data, then specify a FIML approach (i.e., adding *missing = 'fiml'*) in my *lavaan* code. Even though the text-book examples we work have complete data, I will try to include this code so that it will be readily available for you, should you use the as templates for your own data.

In this portion of the ReCentering Psych Stats series we are headed toward more complex models that include both mediation and moderation. Hayes [Hayes, 2018] would call this “conditional process analysis.” Others would simply refer to it as “path analysis.” Although all these terms are sometimes overlapping, *path analysis* is a distinction from *structural equation modeling* (SEM) where latent variables are composed of the observed variables. Let’s take a look at some of the nuances of the whole SEM world and how it relates to PROCESS.

**SEM** is broad term (that could include CFA and path analysis) but is mostly reserved for models with some type of latent variable (i.e., some might exclude path analysis from its definitions). SEM typically uses some form of MLE (not ordinary least squares).

*Latent variables* (circles in the model, below) are those that are “created” in the analytic process but will never appear as a column in your dataset. It may be easiest to think of a latent variable as a scale score – where you sum (or average) the indicator item values to get the score (except we don’t do that). Rather, the LV is “indicated” by variance the indicator/observed/manifest variables share with each other.

The image below is of a simple mediation model but the variables in the model are latent, and indicated by each of the 3 observed/manifest variables. PROCESS (in SPSS) could not assess this model because PROCESS uses ordinary least squares regression and SEM will use a maximum likelihood estimator.

**Confirmatory factor analysis** (CFA) is what we’ll do in psychometrics. Purely SEM, CFA is used to evaluate the structural validity of a scale or measure. In pure CFA, first-order factors represent subscales and a second-order factor (not required) might provide support for a total scale score. For example, in the above figure, the three squares represent the observed (or manifest) items to which a person respond. In CFA, we evaluate their adequacy to represent the latent

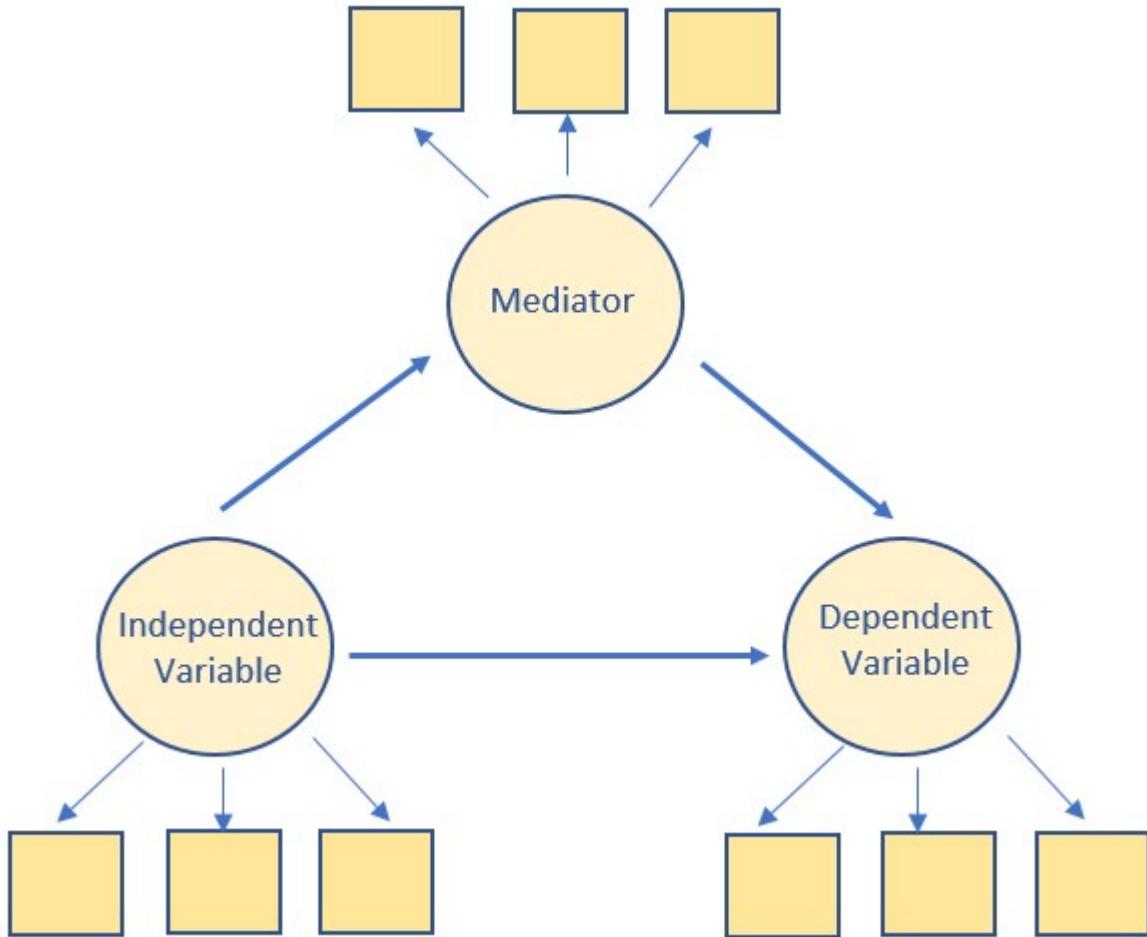


Figure 7.1: Image of a simple mediation model with latent variables

variable (circle) construct. It's a little more complicated than this, but this will get you started. Mediation/indirect effects are not assessed in a pure CFA.

**Path analysis** is a form of SEM, but without latent variables. That is, all the variables in the model are directly observed. They are represented by squares/rectangles and each has a corresponding column in a dataset. PROCESS in SPSS is entirely path analysis.

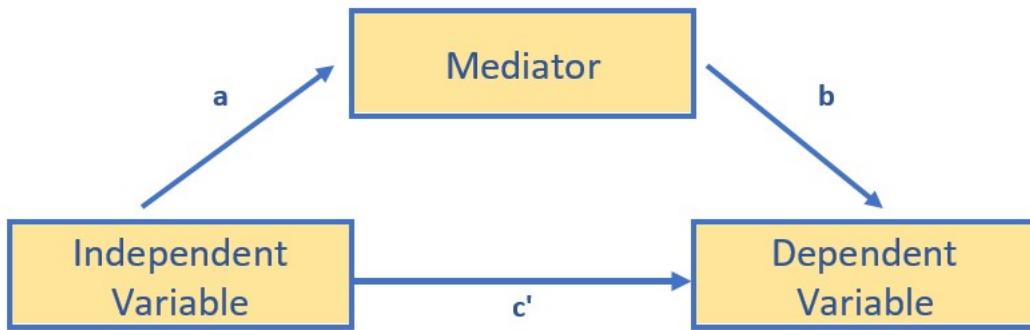


Figure 7.2: Image of a simple mediation in path analysis

**Hybrid models** are a form of SEM that include observed/manifest variables as predictors along with other latent variables. In the diagram below, you see tiny little measurement models (3 indicators that “create” or “inform” an LV, think baby CFA) and one predictor that is manifest. An example might be a categorical predictor (e.g., treatment, control).

## 7.5 Picking up with Moderation

**Moderation:** The effect of X (IV) on some variable Y (DV) is moderated if its size, sign, or strength depends on or can be predicted by W (moderator). In that case, W is said to be a *moderator* of X’s effect on Y. Or, that W and X *interact* in their influence on Y.

Identifying a moderator of an effect helps establish the *boundary conditions* of that effect or the circumstances, stimuli, or type of people for which the effect is large versus small, present versus absent, positive versus negative, and so forth.

**Conditional vs Unconditional Effects:** Consider the following two equations:

$$\hat{Y} = i_y + b_1X + b_2W + e_y$$

and

$$\hat{Y} = i_y + b_1X + b_2W + b_3XW + e_y$$

The first equation constrains X’s effect to be unconditional on W, meaning that it is invariant across all values of W. By introducing the interaction term ( $b_3XW$ ), we can evaluate a model where X’s effect can be dependent on W. That is, for different values of W, X’s effect on Y is different. The resulting equation (#2) is the *simple linear moderation model*. In it, X’s effect on Y is *conditional*.

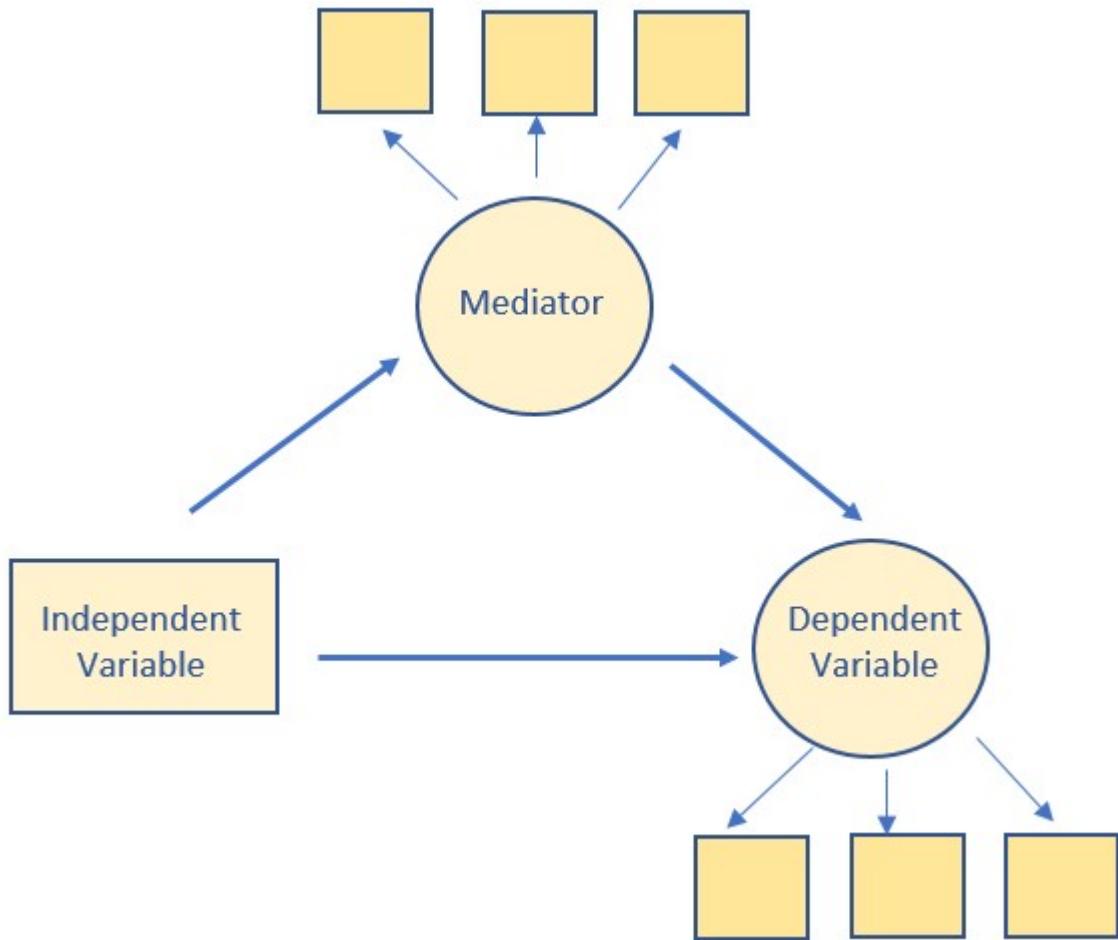


Figure 7.3: Image of a simple mediation in path analysis

## 7.6 Workflow for a Simple Moderation

Below is a workflow comparing the approaches to analyzing a regression model (moderators only) with OLS and MLE.

### Workflow (and comparison) of moderated models with OLS and MLE approaches

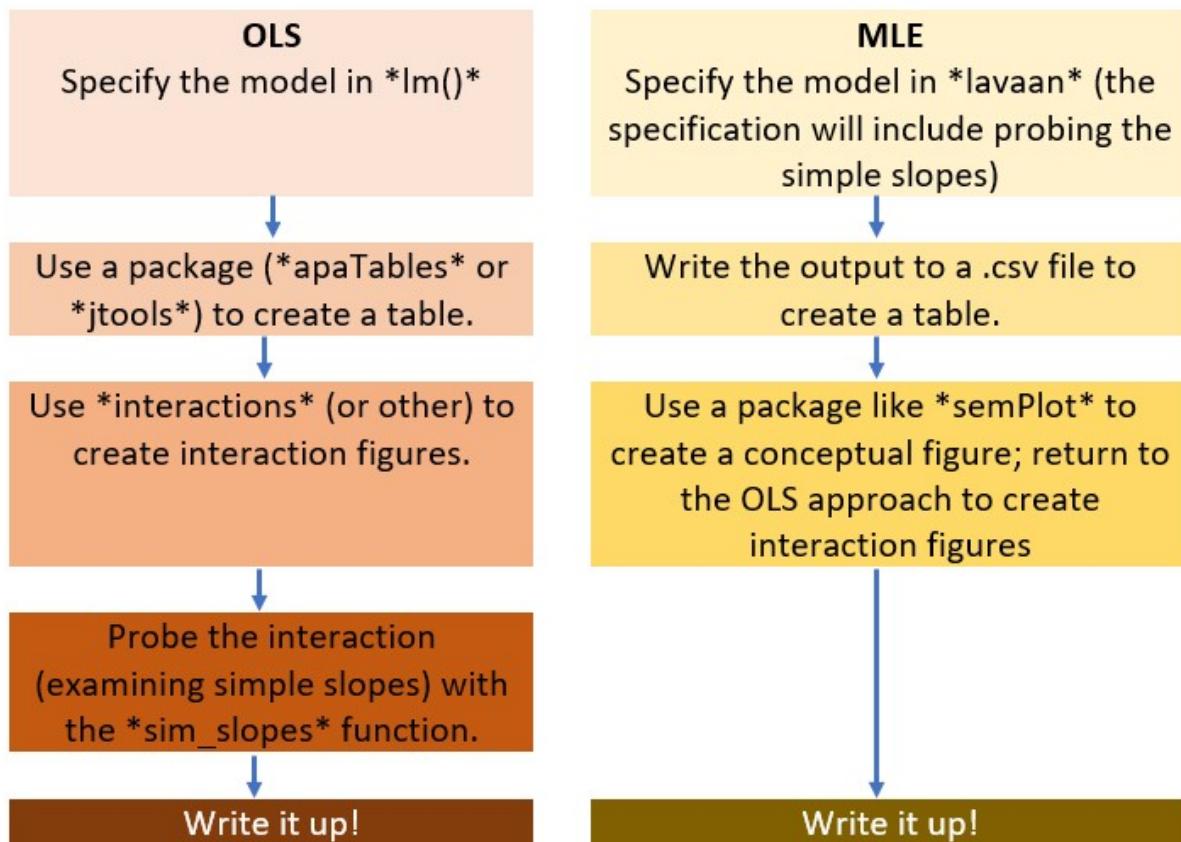


Figure 7.4: Image of a simple mediation in path analysis

The Bonus Track at the end of the chapter includes script templates with just X and Y variables.

## 7.7 Research Vignette

The research vignette comes from the Kim, Kendall, and Cheon's [2017], "Racial Microaggressions, Cultural Mistrust, and Mental Health Outcomes Among Asian American College Students." Participants were 156 Asian American undergraduate students in the Pacific Northwest. The researchers posited the a priori hypothesis that cultural mistrust would mediate the relationship between racial microaggressions and two sets of outcomes: mental health (e.g., depression, anxiety, well-being) and help-seeking.

Variables used in the study included:

- **REMS:** Racial and Ethnic Microaggressions Scale (Nadal, 2011). The scale includes 45 items on a 2-point scale where 0 indicates no experience of a microaggressive event and 1 indicates it was experienced at least once within the past six months. Higher scores indicate more experience of microaggressions.
- **CMI:** Cultural Mistrust Inventory (Terrell & Terrell, 1981). This scale was adapted to assess cultural mistrust harbored among Asian Americans toward individuals from the mainstream U.S. culture (e.g., Whites). The CMI includes 47 items on a 7-point scale where higher scores indicate a higher degree of cultural mistrust.
- **ANX, DEP, PWB:** Subscales of the Mental Health Inventory (Veit & Ware, 1983) that assess the mental health outcomes of anxiety (9 items), depression (4 items), and psychological well-being (14 items). Higher scores (on a 6 point scale) indicate stronger endorsement of the mental health outcome being assessed.
- **HlpSkg:** The Attitudes Toward Seeking Professional Psychological Help – Short Form (Fischer & Farina, 1995) includes 10 items on a 4-point scale (0 = disagree, 3 = agree) where higher scores indicate more favorable attitudes toward help seeking.

### 7.7.1 Simulate Data from the Journal Article

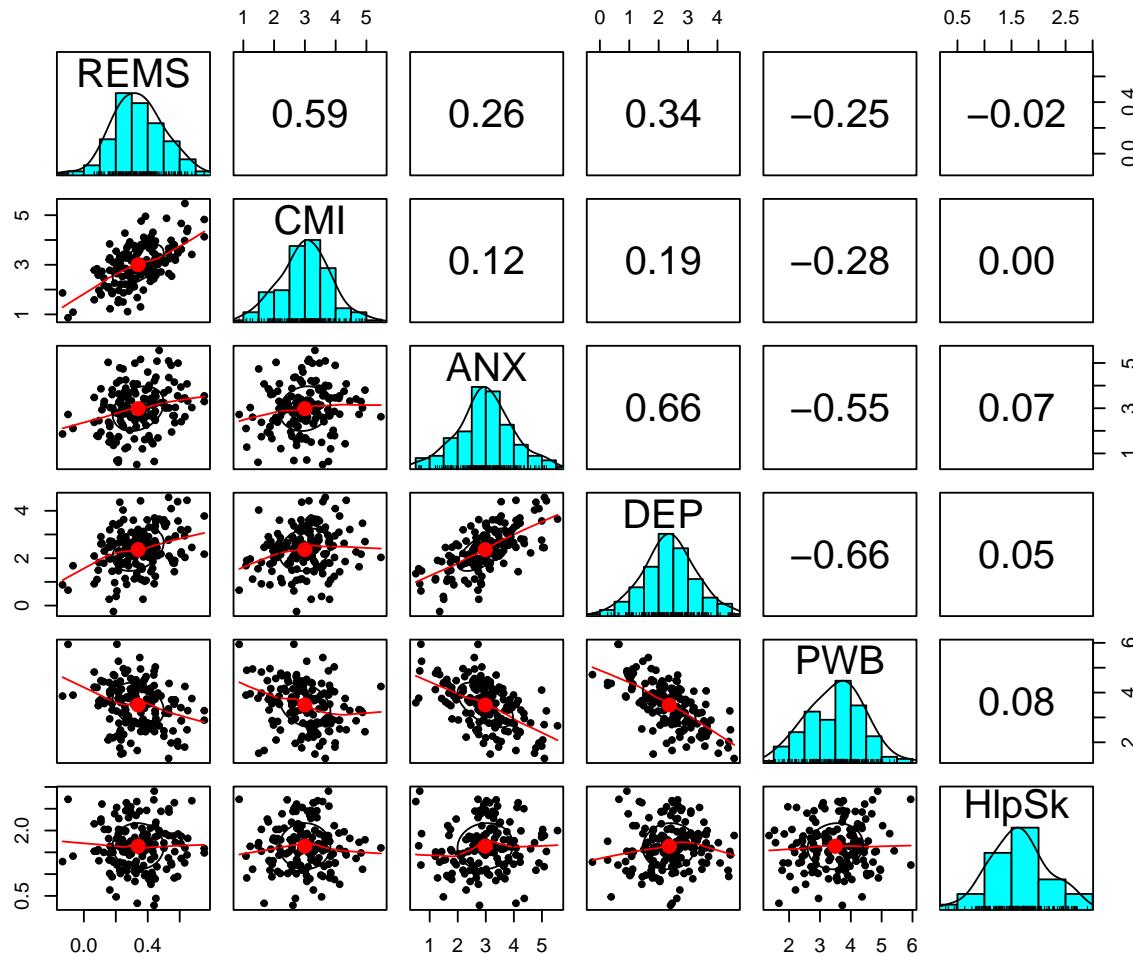
First, we simulate the data from the means, standard deviations, and correlation matrix from the journal article.

```
# Entering the intercorrelations, means, and standard deviations from
# the journal article
mu <- c(0.34, 3, 2.98, 2.36, 3.5, 1.64)
sd <- c(0.16, 0.83, 0.99, 0.9, 0.9, 0.53)
r_mat <- matrix(c(1, 0.59, 0.26, 0.34, -0.25, -0.02, 0.59, 1, 0.12, 0.19,
-0.28, 0, 0.26, 0.12, 1, 0.66, -0.55, 0.07, 0.34, 0.19, 0.66, 1, -0.66,
0.05, -0.25, -0.28, -0.55, -0.66, 1, 0.08, -0.02, 0, 0.07, 0.05, 0.08,
1), ncol = 6)
# Creating a covariance matrix
cov_mat <- sd %*% t(sd) * r_mat

# Set random seed so that the following matrix always gets the same
# results.
set.seed(210409)
library(MASS)
Kim_df <- mvrnorm(n = 156, mu = mu, Sigma = cov_mat, empirical = TRUE)
# renaming the variables
as.data.frame(Kim_df, row.names = NULL, optional = FALSE, make.names = TRUE)
library(tidyverse)
Kim_df <- Kim_df %>%
  as.data.frame %>%
  rename(REMS = V1, CMI = V2, ANX = V3, DEP = V4, PWB = V5, HlpSk = V6)
# Checking our work against the original correlation matrix
# round(cor(Kim_df),3)
```

We can perform a quick check of our data to check its alignment with the journal article, and also get a sense of the bivariate relations with a couple of useful tools. The package *apaTables* produces a journal-ready table with means, standard deviations, and the correlation matrix.

```
library(psych)
psych::pairs.panels(Kim_df)
```



Kim et al. [2017] did not conduct any moderation analyses in their article. Core to their analysis was predicting mental health outcomes (e.g., anxiety, depression, psychological well-being). Their predictors were racial/ethnic microaggressions, cultural mistrust, and help-seeking behaviors. In the majority of their models, REMS was the independent variable, predicting one of the mental health outcomes, mediated by cultural mistrust. Given the strong correlation with REMS ( $r = 0.59$ ) the choice of CMI as a mediator is sound.

In looking at the data, I will ask the question, “Does help-seeking (HlpSk) moderate the relationship between REMS and ANX?”

Here is the formulaic rendering:

$$Y = i_Y + b_1 X + b_2 W + b_3 XW + e_Y$$

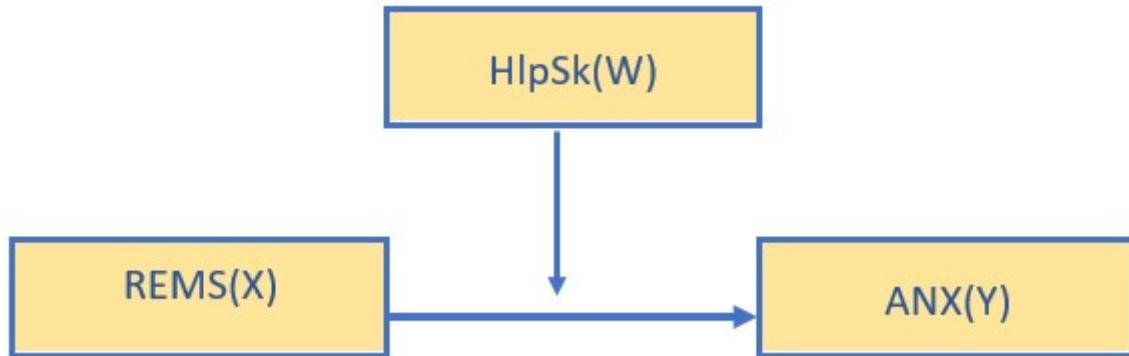


Figure 7.5: Conceptual diagram of a proposed simple moderation model using Kim et al. data

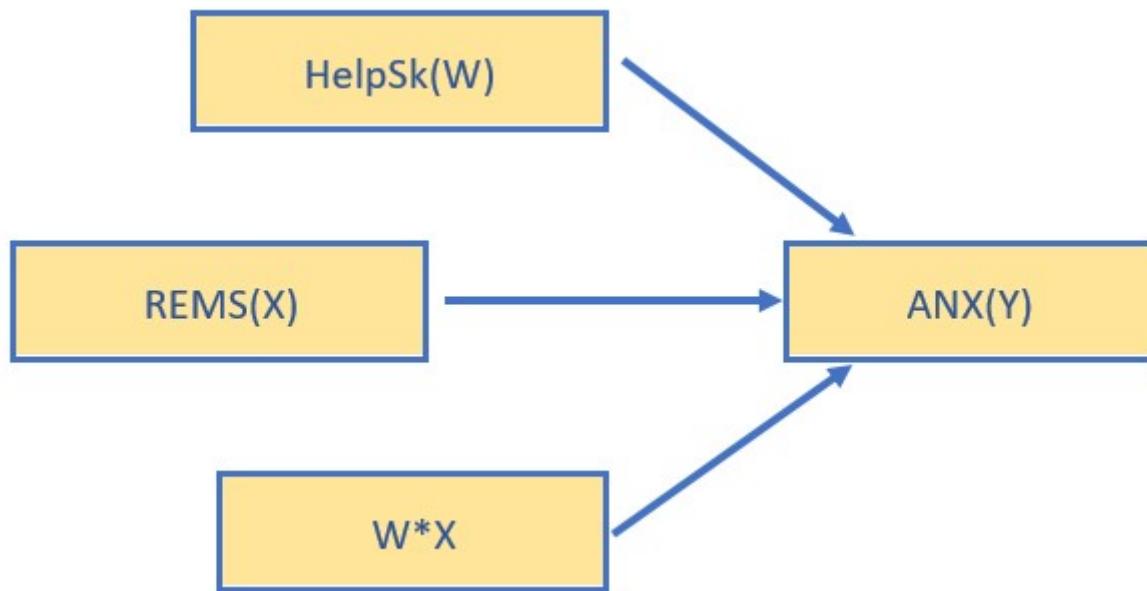


Figure 7.6: Statistical diagram of a proposed simple moderation model using Kim et al. data

## 7.8 Working the Simple Moderation with OLS and MLE

### 7.8.1 OLS with *lm()*

In this demonstration we will use the *lm()* function in base R to evaluate help seeking behaviors (HlpSK) as a moderator to the relationship between racial/ethnic microaggressions (REMS) on anxiety (ANX). Ordinary least squares is the estimator used in *lm()*. We will probe the moderating effect with both pick-a-point and Johnson-Neyman approaches.

Let's specify this simple moderation model with base R's *lm()* function. We'll use the *jtools* package so we get that great *summ* function and *interactions* for an awesome plot.

```
library(jtools) #the summ function creates a terrific regression table
library(interactions)
library(ggplot2)

KimSimpMod <- lm(ANX ~ REMS * HlpSk, data = Kim_df)
# summary(KimSimpMod)
```

**Table 3**

```
KimSimpMod_summ <- summ(KimSimpMod, digits = 3)
KimSimpMod_summ
```

Observations	156
Dependent variable	ANX
Type	OLS linear regression

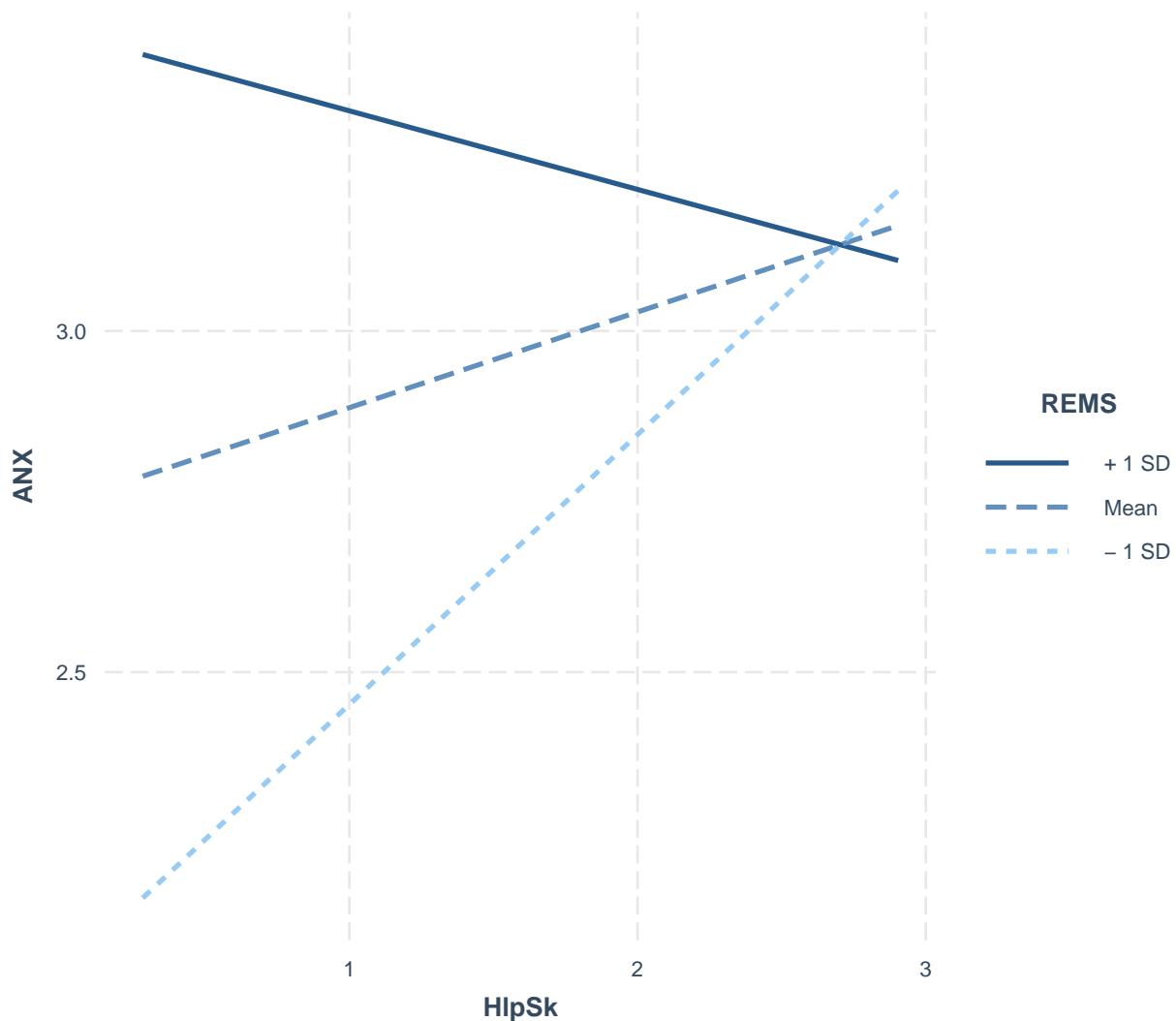
F(3,152)	5.047
R <sup>2</sup>	0.091
Adj. R <sup>2</sup>	0.073

	Est.	S.E.	t val.	p
(Intercept)	1.280	0.618	2.073	0.040
REMS	4.315	1.655	2.607	0.010
HlpSk	0.683	0.350	1.952	0.053
REMS:HlpSk	-1.596	0.938	-1.702	0.091

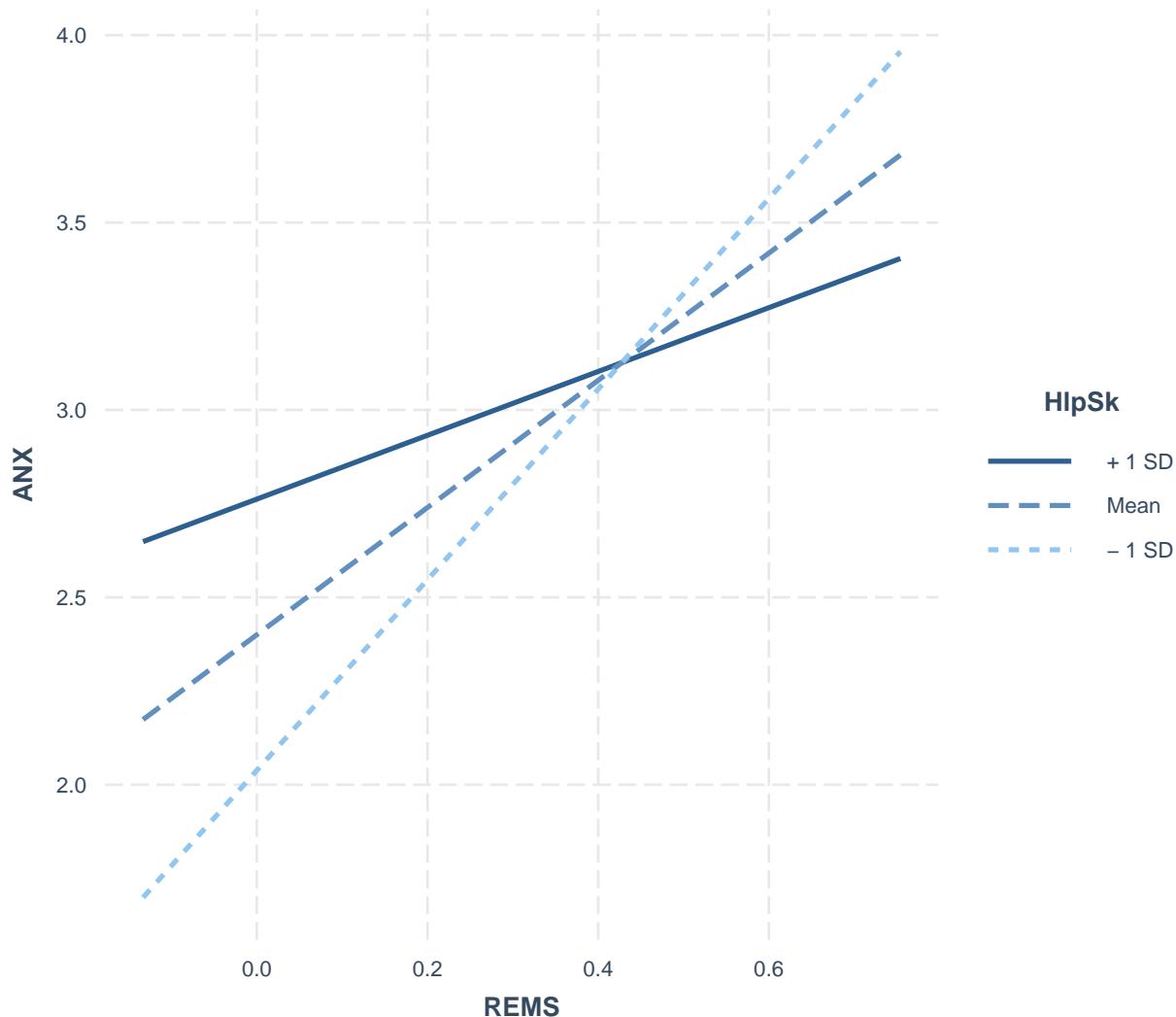
Standard errors: OLS

Looking at these results we can see that the predictors account for about 10% of variance in anxiety. It appears that there is a statistically significant interaction of REMS and HlpSk on Anxiety. The *interaction\_plot()* function from the package, *interactions* can make helpful illustrations. In the case of interactions/moderations, I like to run them “both ways” to see which makes more sense.

```
interact_plot(KimSimpMod, pred = HlpSk, modx = REMS)
```



```
interact_plot(KimSimpMod, pred = REMS, modx = HlpSk)
```



The first figure (where REMS is the moderator) illustrates that for those with the highest experience of racial/ethnic microaggression, the relationship between help-seeking and anxiety is strong and positive. Anxiety is the highest for those who are +1SD above the mean on REMS and who have sought help. This slope is far less strong for those at the mean on REMS and the slope trends negative for those at the lowest REMS.

The second figure places HlpSg as the moderator. The results are the same, merely presented differently. Here we see that at all levels of help seeking, there is a positive relationship between REMS and anxiety. The relationship is the sharpest for those who are at +1SD above the mean on help-seeking. That is, the highest levels of anxiety are among those who experience the most racial and ethnic microaggressions and have the most favorable attitudes toward help-seeking.

Next, let's probe the interaction with simple slopes. With these additional inferential tests we can see where in the distribution of the moderator, X has an effect on Y that is different from zero (and where it does not). There are two common approaches.

The Johnson-Neyman is a *floodlight* approach and provides an indication of the places in the distribution of W (moderator) that X has an effect on Y that is different than zero. The pick-a-point is sometimes called the *analysis of simple slopes* or a *spotlight* approach, probes the distribution at

specific values (often the  $M \pm 1SD$ ).

```
sim_slopes(KimSimpMod, pred = REMS, modx = HlpSk)
```

#### JOHNSON-NEYMAN INTERVAL

When HlpSk is INSIDE the interval [-4.47, 2.01], the slope of REMS is  $p < .05$ .

Note: The range of observed values of HlpSk is [0.28, 2.90]

#### SIMPLE SLOPES ANALYSIS

Slope of REMS when HlpSk = 1.11 (- 1 SD):

Est.	S.E.	t val.	p
2.54	0.72	3.51	0.00

Slope of REMS when HlpSk = 1.64 (Mean):

Est.	S.E.	t val.	p
1.70	0.48	3.53	0.00

Slope of REMS when HlpSk = 2.17 (+ 1 SD):

Est.	S.E.	t val.	p
0.85	0.66	1.30	0.20

```
# sim_slopes(KimSimpMod, pred=GRICntlty, modx = GRMS) #sometimes I
# like to look at it in reverse -- like in the plots
```

The Johnson-Neyman suggests that the relationship between REMS and ANX is statistically significant when HlpSk is above 1.34 (the mean of help-seeking is 1.64). We see the same result in the pick-a-point approach where there is a non-significant relationship between REMS and anxiety when help-seeking is 1SD below the mean. However, there is a statistically significant relationship between help-seeking and REMS when help-seeking is at and above the mean.

#### 7.8.1.1 An APA Style Write-up of OLS results

##### Method/Aalytic Strategy

Data were analyzed with an ordinary least squares approach with the base R function (v. 4.0.4), *lm()*. We specified a model predicting anxiety (ANX) from the interacting effects of racial and ethnic microaggressions (REMS) and attitudes toward help-seeking (HlpSk).

## Results

### Preliminary Analyses

- Missing data analyses and managing missing data
- Bivariate correlations, means, SDs
- Distributional characteristics, assumptions, etc.
- Address limitations and concerns

**Primary Analyses** A multiple regression analysis was conducted to predict anxiety from racial and ethnic microaggressions and attitudes toward help-seeking. Results supported a statistically significant interaction effect that accounted for 8% of the variance. Probing the interaction effect with Johnson-Neyman and pick-a-point approaches indicated that the relationship between REMS and anxiety is non-significant when help-seeking is 1SD below the mean, but is significant when help-seeking is at and above the mean. Results are listed in Table 3. As illustrated in Figure 1, the relationship between REMS and anxiety is the sharpest for those who are at +1SD above the mean on help-seeking. That is, the highest levels of anxiety are among those who experience the most racial and ethnic microaggressions and have the most favorable attitudes toward help-seeking.

### 7.8.2 MLE with *lavaan::sem()*

Specifying the path analysis in lavaan

Things to note:

- MLE has an element of random, by setting the seed we tell it where to “start”...so we all get the same answer
- The code below “draws our model.” It opens and close with ’ marks
- “Labels” (e.g., b1, b2) are useful for identifying the paths. But later in path analysis (mediation) we will use them to do some calculations (i.e., multiplying paths to create an indirect effect). In SEM/CFA (latent variable modeling) we can use them to “fix and free” constraints; the asterisk makes them look like interactions, but they are not
- Interactions are created with the colon
- We can use hashtags internal to the code to remind ourselves (or teach others)
- Following specification of the model, we use the lavaan function *sem()* to conduct the estimation
  - adding *missing = ‘fiml’* is the magic we have been waiting for with regard to missing data
  - bootstrapping is an MLE tool that gives us greater power (more later in mediation)
  - the *summary()* and *parameterEstimates()* functions get us the desired output

```
library(lavaan)
set.seed(210501)
KimSimpModMLE <- "
ANX ~ b1*REMS + b2*HlpSk + b3*REMS:HlpSk
#intercept (constant) of ANX
ANX ~ ANX.mean*1
```

```

#mean of W (HlpSk, in this case) for use in simple slopes
HlpSk ~ HlpSk.mean*1
#variance of W (age, in this case) for use in simple slopes
HlpSk ~~HlpSk.var*HlpSk

#simple slopes
SD.below := b1 + b3*(HlpSk.mean - sqrt(HlpSk.var))
mean := b1 + b3*(HlpSk.mean)
SD.above := b1 + b3*(HlpSk.mean + sqrt(HlpSk.var))
"
kMLE_fit <- sem(KimSimpModMLE, data = Kim_df, missing = "fiml", se = "bootstrap",
bootstrap = 1000)

```

Warning in lav\_partable\_vnames(FLAT, "ov.x", warn = TRUE): lavaan WARNING:  
model syntax contains variance/covariance/intercept formulas  
involving (an) exogenous variable(s): [HlpSk]; These variables  
will now be treated as random introducing additional free  
parameters. If you wish to treat those variables as fixed, remove  
these formulas from the model syntax. Otherwise, consider adding  
the fixed.x = FALSE option.

```

k1summary <- summary(kMLE_fit, standardized = TRUE, fit = TRUE, ci = TRUE)
k1ParamEsts <- parameterEstimates(kMLE_fit, boot.ci.type = "bca.simple",
standardized = TRUE)
k1summary

```

lavaan 0.6.16 ended normally after 9 iterations

Estimator	ML
Optimization method	NLMINB
Number of model parameters	7
Number of observations	156
Number of missing patterns	1

Model Test User Model:

Test statistic	275.935
Degrees of freedom	2
P-value (Chi-square)	0.000

Model Test Baseline Model:

Test statistic	290.749
Degrees of freedom	5
P-value	0.000

User Model versus Baseline Model:

Comparative Fit Index (CFI)	0.041
Tucker-Lewis Index (TLI)	-1.397
Robust Comparative Fit Index (CFI)	0.041
Robust Tucker-Lewis Index (TLI)	-1.397

Loglikelihood and Information Criteria:

Loglikelihood user model (H0)	-333.690
Loglikelihood unrestricted model (H1)	-195.722
Akaike (AIC)	681.380
Bayesian (BIC)	702.728
Sample-size adjusted Bayesian (SABIC)	680.571

Root Mean Square Error of Approximation:

RMSEA	0.937
90 Percent confidence interval - lower	0.846
90 Percent confidence interval - upper	1.032
P-value H_0: RMSEA <= 0.050	0.000
P-value H_0: RMSEA >= 0.080	1.000
Robust RMSEA	0.937
90 Percent confidence interval - lower	0.846
90 Percent confidence interval - upper	1.032
P-value H_0: Robust RMSEA <= 0.050	0.000
P-value H_0: Robust RMSEA >= 0.080	1.000

Standardized Root Mean Square Residual:

SRMR	0.179
------	-------

Parameter Estimates:

Standard errors	Bootstrap
Number of requested bootstrap draws	1000
Number of successful bootstrap draws	1000

Regressions:

	Estimate	Std.Err	z-value	P(> z )	ci.lower	ci.upper	
ANX ~							
REMS	(b1)	4.315	1.364	3.163	0.002	1.769	7.333
HlpSk	(b2)	0.683	0.291	2.350	0.019	0.129	1.320
REMS:HlpS	(b3)	-1.596	0.821	-1.944	0.052	-3.543	-0.131

Std.lv Std.all

4.315	0.632
0.683	0.332
-1.596	-0.485

Intercepts:

		Estimate	Std.Err	z-value	P(> z )	ci.lower	ci.upper
.ANX	(ANX.)	1.280	0.498	2.573	0.010	0.240	2.272
HlpSk	(H1S.)	1.640	0.043	38.507	0.000	1.552	1.729
Std.lv	Std.all						
		1.280	1.176				
		1.640	3.104				

Variances:

		Estimate	Std.Err	z-value	P(> z )	ci.lower	ci.upper
HlpSk	(H1S.)	0.279	0.029	9.506	0.000	0.221	0.338
.ANX		0.886	0.108	8.202	0.000	0.666	1.094
Std.lv	Std.all						
		0.279	1.000				
		0.886	0.748				

Defined Parameters:

		Estimate	Std.Err	z-value	P(> z )	ci.lower	ci.upper
SD.below		2.540	0.572	4.441	0.000	1.424	3.719
mean		1.697	0.420	4.044	0.000	0.846	2.490
SD.above		0.854	0.632	1.351	0.177	-0.487	1.994
Std.lv	Std.all						
		2.540	-0.389				
		1.697	-0.875				
		0.854	-1.360				

### k1ParamEsts

	lhs	op	rhs	label	est	se
1	ANX	~	REMS	b1	4.315	1.364
2	ANX	~	HlpSk	b2	0.683	0.291
3	ANX	~	REMS:HlpSk	b3	-1.596	0.821
4	ANX	~1		ANX.mean	1.280	0.498
5	HlpSk	~1		HlpSk.mean	1.640	0.043
6	HlpSk	~~	HlpSk	HlpSk.var	0.279	0.029
7	ANX	~~		ANX	0.886	0.108
8	REMS	~~		REMS	0.025	0.000
9	REMS	~~	REMS:HlpSk		0.042	0.000
10	REMS:HlpSk	~~	REMS:HlpSk		0.110	0.000
11	REMS	~1			0.340	0.000
12	REMS:HlpSk	~1			0.556	0.000

```

13 SD.below := b1+b3*(HlpSk.mean-sqrt(HlpSk.var)) SD.below 2.540 0.572
14 mean := b1+b3*(HlpSk.mean) mean 1.697 0.420
15 SD.above := b1+b3*(HlpSk.mean+sqrt(HlpSk.var)) SD.above 0.854 0.632
      z pvalue ci.lower ci.upper std.lv std.all std.nox
1  3.163  0.002   1.970   7.812  4.315  0.632   3.965
2  2.350  0.019   0.151   1.395  0.683  0.332   0.100
3 -1.944  0.052  -3.821  -0.263 -1.596 -0.485 -1.467
4  2.573  0.010   0.226   2.169  1.280  1.176   1.176
5 38.507  0.000   1.554   1.730  1.640  3.104   3.104
6  9.506  0.000   0.224   0.338  0.279  1.000   1.000
7  8.202  0.000   0.696   1.116  0.886  0.748   0.748
8    NA     NA   0.025   0.025  0.025  1.000   0.025
9    NA     NA   0.042   0.042  0.042  0.803   0.042
10   NA    NA   0.110   0.110  0.110  1.000   0.110
11   NA    NA   0.340   0.340  0.340  2.132   0.340
12   NA    NA   0.556   0.556  0.556  1.680   0.556
13  4.441  0.000   1.498   3.807  2.540 -0.389  0.879
14  4.044  0.000   0.857   2.511  1.697 -0.875 -0.588
15  1.351  0.177  -0.539   1.960  0.854 -1.360 -2.055

```

Recall, this was our formula:

Here is the formulaic rendering:

$$Y = i_Y + b_1X + b_2W + b_3XW + e_Y$$

Looking at our data here's what we've learned:

$$\hat{Y} = 3.26 + (-1.57)X + (-0.52)W + 1.98XW$$

```

library(semPlot)
semPaths(kMLE_fit, #must identify the model you want to map
         what = "est", ##"est" plots the estimates, but keeps it greyscale with no fading
#whatLabels = "stand", ##"stand" changes to standardized values
         layout = 'tree', rotation = 2, #together, puts predictors on left, IVs on right
#layout = 'circle',
         edge.label.cex = 1.00, #font size of parameter values
#edge.color = "black", #overwrites the green/black coloring
         sizeMan=10, #size of squares/observed/"manifest" variables
         fade=FALSE, #if TRUE, there lines are faded such that weaker lines correspond with lower
         esize=2,
         asize=3,
#label.prop = .5,
         label.font = 2.5, #controls size (I think) of font for labels
         label.scale = TRUE, #if false, the labels will not scale to fit inside the nodes
         nDigits = 3, #decimal places (default is 2)
         residuals = FALSE, #excludes residuals (and variances) from the path diagram
         nCharNodes = 0, #specifies how many characters to abbreviate variable lables; default

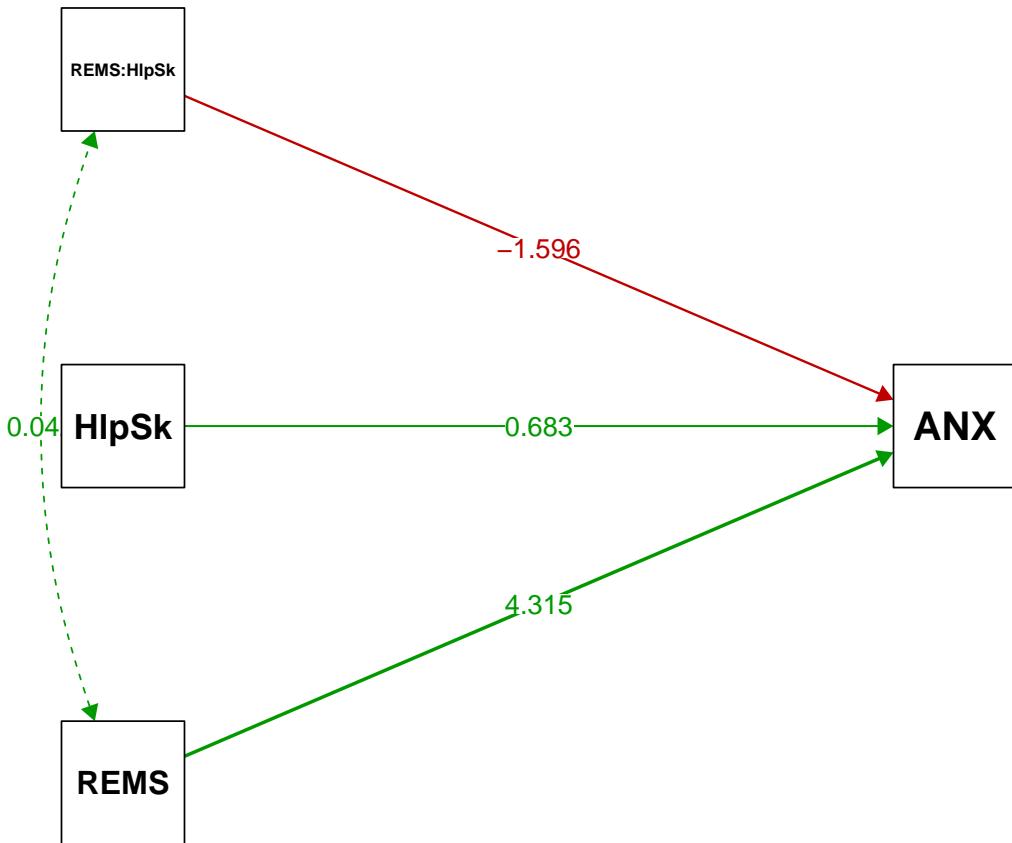
```

```

    intercepts = FALSE, #gets rid of those annoying triangles (intercepts) in the path diagram
)
title("Help Seeking as a Moderator in the Relationship between REMS and ANX")

```

## Help Seeking as a Moderator in the Relationship between REMS and ANX



If I had just run this with lavaan, I would want to plot the interaction and would do so with the OLS methods I demonstrated above.

Attaching package: 'formattable'

The following object is masked from 'package:MASS':

area

### 7.8.3 Tabling the data

In this table, I gather the output from both the OLS and MLE approaches. Youll notice below that the  $B$  weights are identical to the third decimal place (shown). The standard errors and  $p$  values wiggle around a bit, but are consistent with each other (and lead to the same significant/non-significant conclusion). The  $R^2$  values are quite divergent.

Further comparison shows that the OLS output provides an  $F$  statistic that indicates whether or not the overall model is significant. These are commonly reported in Results. In contrast, the MLE output has a page or more of *fit statistics* (e.g., CFI, RMSEA, Chi-square goodness of fit) that are commonly reported in latent variable modeling such as SEM and CFA. Although some researchers will report them in path analysis, I tend to preer the focus on the strength and significance of the regression weights.

Table 4

---

#### A Comparison of OLS and MLE Regression Results

---

	OLS with the <i>lm()</i> in base R			MLE with <i>lavaan</i>		
	<i>B</i>	<i>SE</i>	<i>p</i>	<i>B</i>	<i>SE</i>	<i>p</i>
ANX	1.280	0.618	0.040	1.280	0.498	0.010
(Intercept)						
REMS (X)	4.315	1.655	0.010	4.315	1.364	0.002
HlpSk (W)	0.683	0.350	0.053	0.683	0.291	0.019
REMS:HlpSK (XY)	-1.596	0.938	0.091	-1.596	0.821	0.052
<i>R</i> <sup>2</sup>			<i>R</i> <sup>2</sup>			
9.06%						

---

### 7.8.4 APA Style Writeup

#### Method/Aalytic Strategy

Data were analyzed with a maximum likelihood approach the package, *lavaan* (v. 0.6-7) We specified a model predicting anxiety (ANX) from the interacting effects of racial and ethnic microaggressions (REMS) and attitudes toward help-seeking (HlpSk).

#### Results

##### Preliminary Analyses

- Missing data analyses and managing missing data

- Bivariate correlations, means, SDs
- Distributional characteristics, assumptions, etc.
- Address limitations and concerns

**Primary Analyses** A multiple regression analysis was conducted to predict anxiety from racial and ethnic microaggressions and attitudes toward help-seeking. Results supported a statistically significant interaction effect that accounted for of the variance. Probing the interaction effect with the pick-a-point approaches indicated that the relationship between REMS and anxiety is non-significant when help-seeking is 1SD below the mean, but is significant when help-seeking is at and above the mean. Results are listed in Table 4. As illustrated in Figure 1, the relationship between REMS and anxiety is the sharpest for those who are at +1SD above the mean on help-seeking. That is, the highest levels of anxiety are among those who experience the most racial and ethnic microaggressions and have the most favorable attitudes toward help-seeking.

## 7.9 Residual and Related Questions...

Wait. Why did we do this? And which would you use when?

- As we transition from NHST to statistical modeling we also (generally) transition between OLS and MLE.
- I would use OLS with
  - smaller sample sizes
  - straightforward regression models (linear, multiple, simultaneous, hierarchical)
- I would use MLE with
  - nonlinear models
  - models involving latent variables
  - models with indirect effects
  - (larger sample sizes is prerequisite)

## 7.10 Practice Problems

The suggested practice problem for this chapter is to conduct a simple moderation with both the OLS/*lm()* approach and the MLE/*lavaan* approach.

### 7.10.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed

If this topic feels a bit overwhelming, simply change the random seed in the data simulation, then rework the problem. This should provide minor changes to the data (maybe in the second or third decimal point), but the results will likely be very similar.

---

Assignment Component

1. Assign each variable to the X, Y, and W roles (ok but not required to include a cov)	5	_____
2. Specify and run the OLS/ <i>lm()</i> model	5	_____
3. Probe the interaction with the pick-a-point and Johnson-Neyman approaches	5	_____
4. Create an interaction figure	5	_____
5. Create a table (a package-produced table is fine)	5	_____
6. Create an APA style write-up of the results	5	_____
7. Repeat the analysis in <i>lavaan</i> (specify the model to include probing the interaction)	5	_____
8. Create a model figure.	5	_____
9. Create a table.	5	_____
10. Represent your work in an APA-style write-up	5	_____
11. Explanation to grader	5	_____

---

**7.10.2 Problem #2: Rework the research vignette, but swap one or more variables**

Use the simulated data, but swap out at least one of the variables in the model to conduct the simple moderation using both approaches.

---

Assignment Component

1. Assign each variable to the X, Y, and W roles (ok but not required to include a cov)	5	_____
2. Specify and run the OLS/ <i>lm()</i> model	5	_____
3. Probe the interaction with the pick-a-point and Johnson-Neyman approaches	5	_____
4. Create an interaction figure	5	_____
5. Create a table (a package-produced table is fine)	5	_____
6. Create an APA style write-up of the results	5	_____
7. Repeat the analysis in <i>lavaan</i> (specify the model to include probing the interaction)	5	_____
8. Create a model figure.	5	_____
9. Create a table.	5	_____
10. Represent your work in an APA-style write-up	5	_____
11. Explanation to grader	5	_____

---

**7.10.3 Problem #3: Use other data that is available to you**

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete the simple moderation with both approaches.

---

Assignment Component

---

1. Assign each variable to the X, Y, and W roles (ok but not required to include a cov)	5	_____
2. Specify and run the OLS/ <i>lm()</i> model	5	_____
3. Probe the interaction with the pick-a-point and Johnson-Neyman approaches	5	_____
4. Create an interaction figure	5	_____
5. Create a table (a package-produced table is fine)	5	_____
6. Create an APA style write-up of the results	5	_____
7. Repeat the analysis in <i>lavaan</i> (specify the model to include probing the interaction)	5	_____
8. Create a model figure.	5	_____
9. Create a table.	5	_____
10. Represent your work in an APA-style write-up	5	_____
11. Explanation to grader	5	_____

---

## 7.11 Bonus Track:

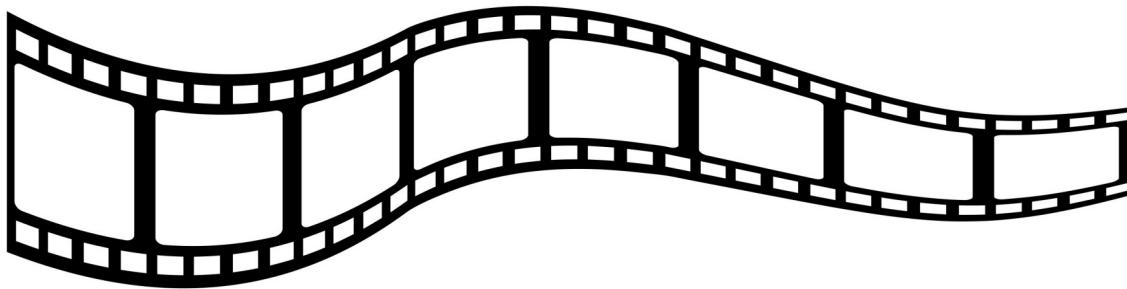


Figure 7.7: Image of a filmstrip

Below is template for a simple moderation conducted with the OLS approach using the base R function, *lm()*

```
library(jtools) #the summ function creates a terrific regression table
library(interactions)
library(ggplot2)

#The regression
#OLSmodel <- lm(Y~X*W, data=my_df)
#summary(KimSimpMod)

#Cool Table
#summ(KimSimpMod, digits = 3)
```

```
#Probe Simple Slopes
#sim_slopes(OLSmodel, pred = X, modx = W)

#Figures
#interact_plot(OLSmodel, pred = W, modx = X)
#interact_plot(OLSmodel, pred = X, modx = W)
```

Below is a template for a simple moderation conducted with the MLE approach using the package, *lavaan*.

```
library(lavaan)
# set.seed(210501) MLEmodel <- ' Y ~ b1*X + b2*W + b3*X:W intercept
# (constant) of Y Y ~ Y.mean*1 mean of W for use in simple slopes W ~
# W.mean*1 variance of W for use in simple slopes W ~ W.var*W

# simple slopes SD.below := b1 + b3*(W.mean - sqrt(W.var)) mean := b1
# + b3*(W.mean) SD.above := b1 + b3*(W.mean + sqrt(W.var))
#
# MLEmod_fit <- semMLEmodel, data = my_df, missing = 'fiml', se =
# 'bootstrap', bootstrap = 1000) MLEmod_fit_summary <-
# summary(MLEmod_fit, standardized = TRUE, rsq=T, ci=TRUE)
# MLEmodParamEsts <- parameterEstimates(MLEmod_fit, boot.ci.type =
# 'bca.simple', standardized=TRUE) MLEmod_fit_summary MLEmodParamEsts
```

```
# library(semPlot) semPathsMLEmod_fit, #must identify the model you
# want to map what = 'est', #'est' plots the estimates, but keeps it
# greyscale with no fading whatLabels = 'stand', #'stand' changes to
# standardized values layout = 'tree', rotation = 2, #together, puts
# predictors on left, IVs on right layout = 'circle', edge.label.cex
# = 1.00, #font size of parameter values edge.color = 'black',
# #overwrites the green/black coloring sizeMan=10, #size of
# squares/observed/'manifest' variables fade=FALSE, #if TRUE, there
# lines are faded such that weaker lines correspond with lower values
# -- a cool effect, but tough for journals esize=2, asize=3,
# label.prop = .5, label.font = 2.5, #controls size (I think) of font
# for labels label.scale = TRUE, #if false, the labels will not scale
# to fit inside the nodes nDigits = 3, #decimal places (default is 2)
# residuals = FALSE, #excludes residuals (and variances) from the path
# diagram nCharNodes = 0, #specifies how many characters to
# abbreviate variable lables; default is 3. If 0, uses your entire
# variable label and adjusts fontsize (which could be a downside)
# intercepts = FALSE, #gets rid of those annoying triangles
# (intercepts) in the path diagram) ) title('Help Seeking as a
# Moderator in the Relationship between REMS and ANX')
```



# **CONDITIONAL PROCESS ANALYSIS**



# Chapter 8

## Moderated Mediation

### [Screencasted Lecture Link](#)

The focus of this lecture is the moderated mediation. That is, are the effects of the indirect effect (sign, significance, strength, presence/absence) *conditional* on the effects of the moderator.

At the outset, please note that although I rely heavily on Hayes [2018] text and materials, I am using the R package *lavaan* in these chapters. Very recently, Hayes has introduced a [PROCESS macro for R](#). Because I am not yet up-to-speed on using this macro (it is not a typical R package) and because we will use *lavaan* for confirmatory factor analysis and structural equation modeling, I have chosen to utilize the *lavaan* package. A substantial difference is that the PROCESS macros use ordinary least squares and *lavaan* uses maximum likelihood estimators.

### 8.1 Navigating this Lesson

There is about 1 hour and 15 minutes of lecture. If you work through the materials with me it would be plan for an additional hour and a half.

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

#### 8.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Outline a process of evaluating a moderated mediation in a piecewise [Hayes, 2018] approach to model building
- Recognize conditional process modeling from R script.
- Using the R package *lavaan*,
  - specify a model with indirect effects,

- identify and interpret B weights,  $p$  values, and *CIs* for total, direct, and indirect effects,
- calculate the total effects of X and M on Y,
- identify the proportion of variance accounted for in predicting M and Y.
- Regarding conditional indirect effects
  - Interpret an index of moderated mediation
  - Know the essential components of calculating an index of moderated mediation
  - Probe a conditional indirect effect
- Interpret “the usual” things we find in regression: B/beta weights, R,  $R^2$ , and figures

### 8.1.2 Planning for Practice

The suggestions for homework are graded in complexity and, if you like, can extend from the prior chapter on simple moderation. If you choose the first or second options, you can further amend the simulated data by making further variations such as sample size.

- Rework the problem in the chapter by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- There are a number of variables in the dataset. Swap out one or more variables in the moderated mediation and compare your solution to the one in the chapter (and/or one you mimicked in the journal article).
- Conduct a moderated mediation with data to which you have access. This could include data you simulate on your own or from a published article.

### 8.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press. Available as an ebook from the SPU library: <https://ebookcentral-proquest-com.ezproxy.spu.edu/lib/spu/detail.action?docID=5109647>
  - **Chapter 11, CPA fundamentals:** In this chapter Hayes disentangles conditional indirect effects.
  - **Chapter 12, More CPA examples:** Among the examples is one that includes covariates.
  - **Appendix A: Using Process:** An essential tool for PROCESS users because, even when we are in the R environment, this is the “idea book.” That is, the place where all the path models are presented in figures.

- Lewis, J. A., Williams, M. G., Peppers, E. J., & Gadson, C. A. (2017). Applying intersectionality to explore the relations between gendered racism and health among Black women. *Journal of Counseling Psychology*, 64(5), 475–486. <https://doi-org.ezproxy.spu.edu/10.1037/cou0000231>

### 8.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
if (!require(lavaan)) {
    install.packages("lavaan")
}

## Loading required package: lavaan

## This is lavaan 0.6-16
## lavaan is FREE software! Please report any bugs.

if (!require(semPlot)) {
    install.packages("semPlot")
}

## Loading required package: semPlot

if (!require(tidyverse)) {
    install.packages("tidyverse")
}

## Loading required package: tidyverse

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## vforcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyrr    1.3.0
## v purrr    1.0.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```

if (!require(psych)) {
  install.packages("psych")
}

## Loading required package: psych
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
##
## The following object is masked from 'package:lavaan':
##
##     cor2cov

if (!require(jtools)) {
  install.packages("jtools")
}

## Loading required package: jtools

```

## 8.2 Conditional Process Analysis

### 8.2.1 The definitional and conceptual

Hayes [2018] coined the term and suggests we also talk about “conditional process modeling.”

**Conditional process analysis:** used when the analytical goal is to describe and understand the conditional nature of the mechanism or mechanisms by which a variable transmits its effect on another.

We are integrating moderation and mediation mechanisms together into a single integrated analytical model.

- **Mediator:** Any causal system in which at least one causal antecedent X variable is proposed as influencing an outcome Y through a intervening variable M. In this model, there are two pathways by which X can influence Y: *direct* effect of X on Y, and *indirect* effect of X on Y through M.
  - Answers question, “How does X affect Y”
  - Partitions the X-to-Y relationship into two paths of influence: direct, indirect.
  - Indirect effect contains two components (a,b) that when multiplied ( $a \times b$ ) yield an estimate of how much these two cases that differ by one unit on X are estimated to differ on Y through the effect of X on M, which in turn affects Y.
  - Keywords: how, through, via, indirect effect

- **Moderator:** The effect of X on some variable Y is moderated by W if its size, sign, or strength depends on or can be predicted by W.
  - Stated another way, W and X *interact* in their influence on Y.
  - Moderators help establish the boundary conditions of an effect or the circumstances, stimuli, or type of people for which the effect is large v. small, present v. absent, positive v. negative, and so forth.
  - Keywords: “it depends,” interaction effect.

**Why should we engage both mediators and moderators?** Hayes [2018] suggest that if we have only a mediator(s) in the model that we lose information if we “reduce complex responses that no doubt differ from person to person or situation to situation” (p. 394). He adds that “all effects are moderated by something” (p. 394). Correspondingly, he recommends we add them to a mediation analysis.

Hayes [2018] suggests that “more complete” (p. 395) analyses model the mechanisms at work linking X to Y (mediator[s]) while simultaneously allowing those effects to be contingent on context, circumstance, or individual difference (moderator[s]).

**What are conditional direct and indirect effects?** Mediation analyses produce indirect (the product of a sequence of effects that are assumed to be causal) and direct (the unique contribution of X to Y, controlling for other variables in the model) effects. These effects (the X-to-Y/direct and X-to-M-to-Y/indirect), can also be moderated. This is our quest! Figure 11.2 in Hayes’ text [2018] illustrates conceptually and statistically that we can specify moderation of any combination of direct and indirect paths/effects.

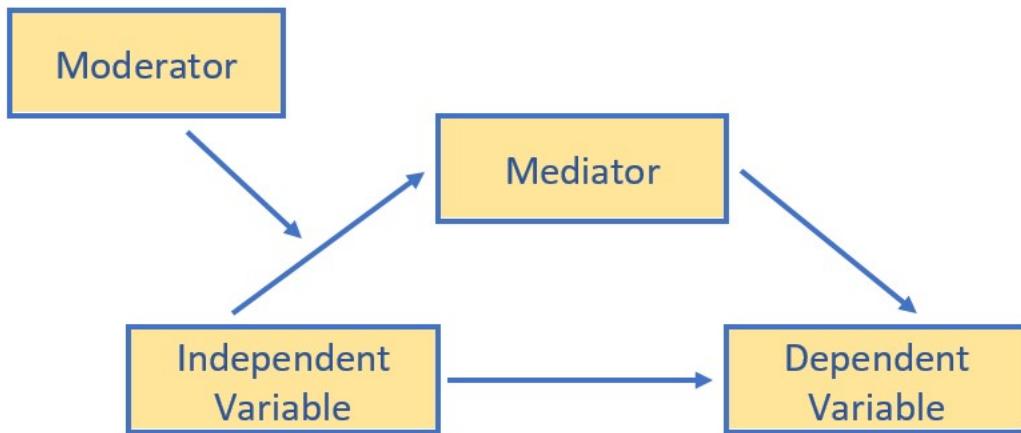


Figure 8.1: Image of conditional process analysis model where the moderator is hypothesized to change the a path; the path between the IV and mediator

Within the CPA framework we have lots of options that generally fall into two categories:

- *Moderated mediation:* when an indirect effect of X on Y through M is moderated; the mechanism represented by the *X-to-M-to-Y* chain of events operates to varying degrees (or not at all) for certain people or in certain contexts.

- Any model in which the indirect effect ( $a^*b$ ) changes as a function of one or more moderators. These moderators can be operating on the  $a$ ,  $b$ , or  $c'$  paths or any possible combination of the three
- $X$  could moderate its own indirect effect on  $Y$  through  $M$  if the effect of  $M$  on  $Y$  depends on  $X$ , or
- The indirect effect of  $X$  on  $Y$  through  $M$  could be contingent on a fourth variable if that fourth variable  $W$  moderates one or more of the relationships in a three-variable causal system, or
- An indirect effect could be contingent on a moderator variable
- *Mediated moderation:* an interaction between  $X$  and some moderator  $W$  on  $Y$  is carried through a mediator  $M$ ;
  - mediated moderation analysis is simply a mediation analysis with the product of two variables serving as the causal agent of focus
  - An interaction between a moderator  $W$  and causal agent  $X$  on outcome  $Y$  could operate through a mediator  $M$

Hayes argues that the mediated moderation hypotheses are “regularly articulated and tested by scientists” [2018, p. 459]. He warns, though, that we should not confuse the “abundance of published examples of mediated moderation analyses...with the meaningfulness of the procedure itself” (p. 460). He later adds that mediation moderation is “neither interesting nor meaningful.” Why?

- Conceptualizing a process in terms of a mediated moderation misdirects attention toward a variable in the model that actually doesn’t measure anything.
- Most often there are moderated mediation models that are identical in equations and resulting coefficients - the difference is in the resulting attentional focus and interpretation.
- Hayes [2018] recommends that models proposing mediated moderation be recast in terms of moderated mediation process.
- Consequently, we will not work a mediated moderation, but there is an example in chapter 12.

### 8.2.2 Hayes’ [2018] Piecewise Approach to Building Models

In summarizing a strategic approach for testing structural equation models, Joreskog [Joreskog, 1993] identified three scenarios:

- *strictly confirmatory:* the traditional NHST approach of proposing a single, theoretically derived, model, and after analyzing the data either rejects or fails to reject the model. No further modifications are made/allowed.
- *alternative models:* the researcher proposes competing (also theoretically derived) models. Following analysis of a single set of empirical data, he or she selects one model as appropriate in representing the sample data.
- *model generating:* A priori, the researcher acknowledges that they may/may not find what they have theoretically proposed. So, a priori, they acknowledge that in the absence of ideal fit (which is the usual circumstance), they will proceed in an exploratory fashion to respecify/re-estimate the model. The goal is to find a model that is both substantively meaningful and statistically well-fitting.

A legacy of our field is the *strictly confirmatory* approach. I am thrilled when I see research experts (e.g., [Byrne, 2016]) openly endorse a model building approach. In Chapter 12, Hayes [2018] demonstrates the piecewise approach to building (and understanding) a complex model.

### 8.3 Workflow for Moderated Mediation

At this point in this OER's development, I don't have a workflow graphic developed for this statistic. However, Hayes' [2018] *piecewise* approach to model testing/building is really the workflow. The secret is to decompose the model into its simplest moderations and mediations and analyze them separately before assembling them. When we get to the model we will analyze with this research vignette, a series of diagrams will make this more clear.

Additionally, at the end of the chapter, I offer a template of R script for the popular moderated mediation (a single moderator influencing both the  $a$  and  $c'$  paths).

### 8.4 Research Vignette

Once again the research vignette comes from the Lewis, Williams, Peppers, and Gadson's [2017] study titled, "Applying Intersectionality to Explore the Relations Between Gendered Racism and Health Among Black Women." The study was published in the Journal of Counseling Psychology. Participants were 231 Black women who completed an online survey.

Variables used in the study included:

- **GRMS:** Gendered Racial Microaggressions Scale [Lewis and Neville, 2015] is a 26-item scale that assesses the frequency of nonverbal, verbal, and behavioral negative racial and gender slights experienced by Black women. Scaling is along six points ranging from 0 (never) to 5 (once a week or more). Higher scores indicate a greater frequency of gendered racial microaggressions. An example item is, "Someone has made a sexually inappropriate comment about my butt, hips, or thighs."
- **MntlHlth** and **PhysHlth**: Short Form Health Survey - Version 2 [Ware et al., 1995] is a 12-item scale used to report self-reported mental (six items) and physical health (six items). Higher scores indicate higher mental health (e.g., little or no psychological distress) and physical health (e.g., little or no reported symptoms in physical functioning). An example of an item assessing mental health was, "How much of the time during the last 4 weeks have you felt calm and peaceful?"; an example of a physical health item was, "During the past 4 weeks, how much did pain interfere with your normal work?"
- **Sprtlty, SocSup, Engmgt**, and **DisEngmgt** are four subscales from the Brief Coping with Problems Experienced Inventory [Carver, 1997]. The 28 items on this scale are presented on a 4-point scale ranging from 1 (*I usually do not do this at all*) to 4(*I usually do this a lot*). Higher scores indicate a respondents' tendency to engage in a particular strategy. Instructions were modified to ask how the female participants responded to recent experiences of racism and sexism as Black women. The four subscales included spirituality (religion, acceptance, planning), interconnectedness/social support (vent emotions, emotional support,instrumental social support), problem-oriented/engagement coping (active coping,

humor, positive reinterpretation/positive reframing), and disengagement coping (behavioral disengagement, substance abuse, denial, self-blame, self-distraction).

- **GRIcntly:** The Multidimensional Inventory of Black Identity Centrality subscale [Sellers et al.] was modified to measure the intersection of racial and gender identity centrality. The scale included 10 items scaled from 1 (*strongly disagree*) to 7 (*strongly agree*). An example item was, “Being a *Black woman* is important to my self-image.” Higher scores indicated higher levels of gendered racial identity centrality.

#### 8.4.1 Simulating the data from the journal article

First, we simulate the data from the means, standard deviations, and correlation matrix from the journal article.

```
# Entering the intercorrelations, means, and standard deviations from
# the journal article
LEWmu <- c(1.99, 2.82, 2.48, 2.32, 1.75, 5.71, 21.37, 21.07)
LEWsd <- c(0.9, 0.7, 0.81, 0.61, 0.53, 1.03, 3.83, 4.66)
LEWr_mat <- matrix(c(1, 0.2, 0.28, 0.3, 0.41, 0.19, -0.32, -0.18, 0.2,
1, 0.49, 0.57, 0.22, 0.13, -0.06, -0.13, 0.28, 0.49, 1, 0.46, 0.26,
0.38, -0.18, -0.08, 0.3, 0.57, 0.46, 1, 0.37, 0.08, -0.14, -0.06, 0.41,
0.22, 0.26, 0.37, 1, 0.05, -0.54, -0.28, 0.19, 0.13, 0.38, 0.08, 0.05,
1, -0.1, 0.14, -0.32, -0.06, -0.18, -0.14, -0.54, -0.1, 1, 0.47, -0.18,
-0.13, -0.08, -0.06, -0.28, 0.14, 0.47, 1), ncol = 8)

# Creating a covariance matrix

LEWcov_mat <- LEWsd %*% t(LEWsd) * LEWr_mat
LEWcov_mat
```

```
##      [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,]  0.81000  0.12600  0.204120  0.164700  0.195570  0.176130 -1.103040
## [2,]  0.12600  0.49000  0.277830  0.243390  0.081620  0.093730 -0.160860
## [3,]  0.20412  0.27783  0.656100  0.227286  0.111618  0.317034 -0.558414
## [4,]  0.16470  0.24339  0.227286  0.372100  0.119621  0.050264 -0.327082
## [5,]  0.19557  0.08162  0.111618  0.119621  0.280900  0.027295 -1.096146
## [6,]  0.17613  0.09373  0.317034  0.050264  0.027295  1.060900 -0.394490
## [7,] -1.10304 -0.16086 -0.558414 -0.327082 -1.096146 -0.394490 14.668900
## [8,] -0.75492 -0.42406 -0.301968 -0.170556 -0.691544  0.671972  8.388466
##      [,8]
## [1,] -0.754920
## [2,] -0.424060
## [3,] -0.301968
## [4,] -0.170556
## [5,] -0.691544
## [6,]  0.671972
## [7,]  8.388466
## [8,] 21.715600
```

```
# Set random seed so that the following matrix always gets the same
# results.
set.seed(210403)
library(MASS)
```

```
##
```

```
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
##      select
```

```
Lewis_df <- mvrnorm(n = 212, mu = LEWmu, Sigma = LEWcov_mat, empirical = TRUE)
colMeans(Lewis_df)
```

```
## [1] 1.99 2.82 2.48 2.32 1.75 5.71 21.37 21.07
```

```
# Checking our work against the original correlation matrix
cor(Lewis_df)
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
## [1,] 1.00  0.20  0.28  0.30  0.41  0.19 -0.32 -0.18
## [2,] 0.20  1.00  0.49  0.57  0.22  0.13 -0.06 -0.13
## [3,] 0.28  0.49  1.00  0.46  0.26  0.38 -0.18 -0.08
## [4,] 0.30  0.57  0.46  1.00  0.37  0.08 -0.14 -0.06
## [5,] 0.41  0.22  0.26  0.37  1.00  0.05 -0.54 -0.28
## [6,] 0.19  0.13  0.38  0.08  0.05  1.00 -0.10  0.14
## [7,] -0.32 -0.06 -0.18 -0.14 -0.54 -0.10  1.00  0.47
## [8,] -0.18 -0.13 -0.08 -0.06 -0.28  0.14  0.47  1.00
```

Rename the variables

```
as.data.frame(Lewis_df, row.names = NULL, optional = FALSE, make.names = TRUE)
library(tidyverse)
Lewis_df <- Lewis_df %>%
  as.data.frame %>%
  rename(GRMS = V1, Sprtlty = V2, SocSup = V3, Engmgt = V4, DisEngmt = V5,
         GRIcntlty = V6, MntlHlth = V7, PhysHlth = V8)
```

```
head(Lewis_df)
```

	GRMS	Sprtlty	SocSup	Engmgt	DisEngmt	GRIcntlty	MntlHlth	PhysHlth
## 1	0.7792361	2.628957	1.758948	1.691459	1.062341	5.533258	22.70042	19.42231
## 2	1.5729406	1.943789	1.101567	2.446707	1.885076	5.806530	22.67086	22.25516
## 3	1.9586843	3.039406	1.591625	2.428866	1.635518	5.166721	19.06958	23.23199
## 4	0.6532324	2.624590	1.039778	1.495290	1.506393	4.276244	23.90836	18.74549
## 5	2.8280150	3.242341	2.202956	1.553723	1.024422	5.730293	22.86224	18.80227
## 6	1.2809196	3.052410	4.097964	2.727955	1.565009	8.474002	19.13631	24.48153

### 8.4.2 Quick peek at the data

```
library(psych)
psych::describe(Lewis_df)

##          vars   n   mean    sd median trimmed   mad   min   max range skew
## GRMS      1 212  1.99  0.90    2.01    2.00  0.93 -0.75  4.24  4.99 -0.12
## Sprtlty   2 212  2.82  0.70    2.75    2.82  0.65  0.46  4.68  4.23 -0.06
## SocSup    3 212  2.48  0.81    2.47    2.46  0.77 -0.32  4.68  5.00  0.11
## Engmgt    4 212  2.32  0.61    2.33    2.32  0.57  0.37  4.08  3.71 -0.02
## DisEngmt  5 212  1.75  0.53    1.75    1.75  0.55  0.58  3.00  2.42 -0.04
## GRICntly  6 212  5.71  1.03    5.67    5.68  1.00  3.08  9.40  6.32  0.32
## MntlHlth  7 212 21.37  3.83   21.60   21.46  4.29 11.65 31.90 20.25 -0.15
## PhysHlth  8 212 21.07  4.66   20.79   21.03  4.68  8.43 33.71 25.28  0.07
##          kurtosis   se
## GRMS        -0.14  0.06
## Sprtlty     0.34  0.05
## SocSup      0.41  0.06
## Engmgt      0.22  0.04
## DisEngmt   -0.64  0.04
## GRICntly   0.36  0.07
## MntlHlth   -0.54  0.26
## PhysHlth   -0.18  0.32
```

And a quick peek at a correlation matrix.

```
library(apaTables)

## Registered S3 methods overwritten by 'broom':
##   method           from
##   tidy.glht       jtools
##   tidy.summary.glht jtools

apa.cor.table(Lewis_df, show.conf.interval = FALSE)

## The ability to suppress reporting of reporting confidence intervals has been deprecated in +
## The function argument show.conf.interval will be removed in a later version.

##
##
## Means, standard deviations, and correlations with confidence intervals
##
##
##      Variable      M      SD      1          2          3          4
```



```
## Note. M and SD are used to represent mean and standard deviation, respectively.
## Values in square brackets indicate the 95% confidence interval.
## The confidence interval is a plausible range of population correlations
## that could have caused the sample correlation (Cumming, 2014).
## * indicates p < .05. ** indicates p < .01.
##
```

## 8.5 Working the Moderated Mediation

The model we are testing is predicting a mental health (MntlHlth, Y) from gendered racial microaggressions (GRMS,X), mediated by disengagement coping (DisEngmt, M). The relationship between gendered racial microaggressions and disengagement coping (i.e., the  $a$  path) is expected to be moderated by gendered racial identity centrality (GRIcntly, W). Gendered racial identity centrality is also expected to moderate the path between gendered racial microaggressions and mental health (i.e., the  $c'$  path). Thus, the specified model involves the evaluation of a conditional indirect effect.

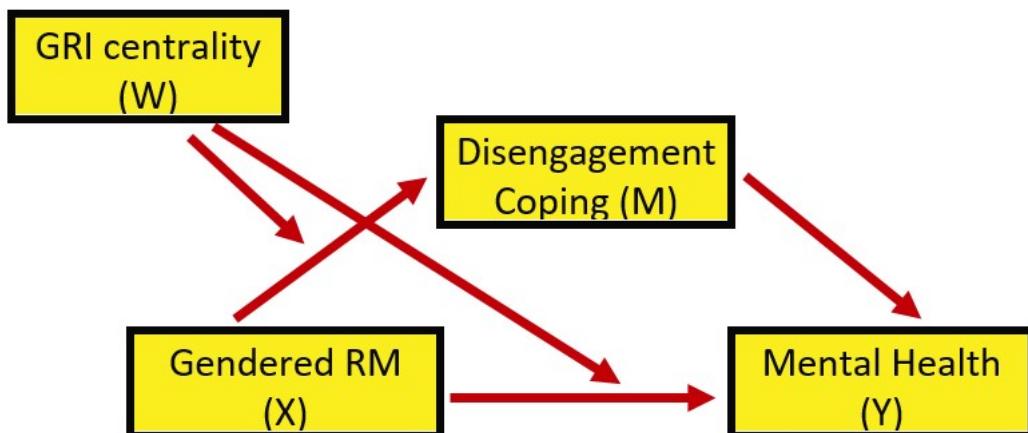


Figure 8.2: Image of conceptual representation of the conditional process analysis model where the moderator is hypothesized to change the  $a$  and  $c'$  paths

Hayes' [2018] textbook and training materials frequently display the conceptual (above) and statistical models (below). These help facilitate understanding.

Looking at the diagram, with two consequent variables (i.e., those with arrows pointing to them) we can see two equations are needed to explain the model:

$$M = i_M + a_1X + a_2W + a_3XW + e_M$$

$$Y = i_Y + c'_1X + c'_2W + c'_3XW + bM + e_Y$$

When we have complicated models such as these, Hayes [2018] suggests a piecewise approach to model building. Specifically, he decompose the model into its aggregate parts: a simple mediation and two simple moderation.

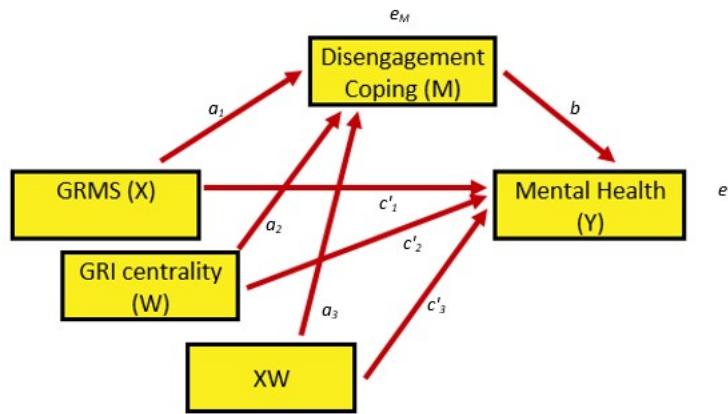


Figure 8.3: Image of statistical representation of the conditional process analysis model where the moderator is hypothesized to change the  $a$  and  $c'$  paths

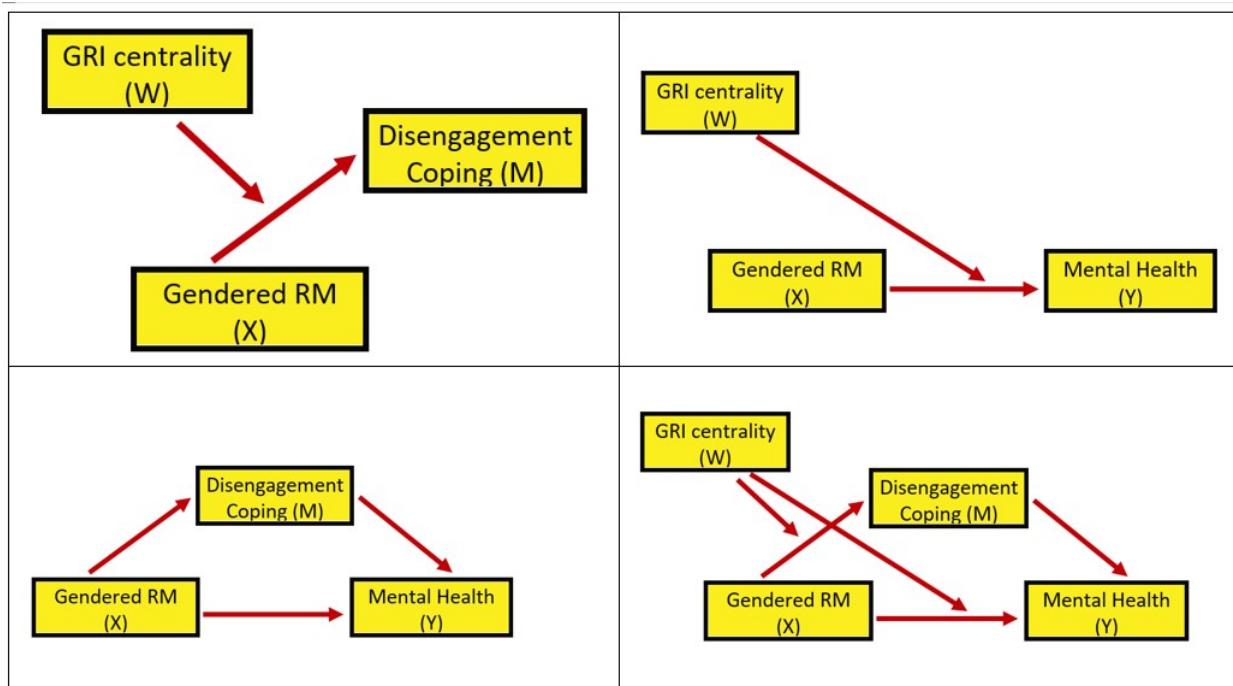


Figure 8.4: Image of statistical representation of the conditional process analysis model where the moderator is hypothesized to change the  $a$  and  $c'$  paths

Let's start with the simple moderations.

### 8.5.1 Piecewise Assembly of the Moderated Mediation

#### 8.5.1.1 Analysis #1: A simple moderation

We are asking, “Does GRI centrality moderate the relationship between gendered racial microaggressions and disengagement coping?

$Y$  = disengagement coping  $X$  = gendered racial microaggressions  $W$  = GRI centrality

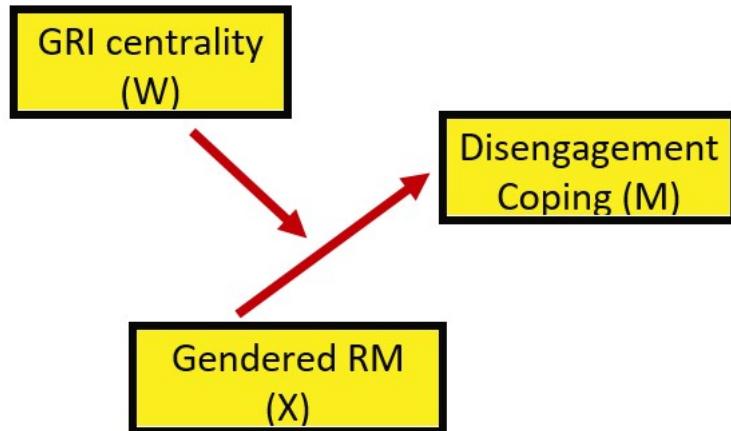


Figure 8.5: Image of statistical representation of the simple moderation estimating DisEngmt from GRMS, moderated by GRIcntly

The formula we are estimating:

$$Y = b_0 + b_1 X + b_2 W + b_3 XW + e_Y$$

Let's specify this simple moderation model with base R's `lm()` function. Let's use the `jtools` package so we get that great `summ` function and `interactions` for the awesome plot.

Since we are just working to understand our moderations, we can run them with “regular old” ordinary least squares.

```

library(jtools) #the summ function creates a terrific regression table
library(interactions)
library(ggplot2)

Mod_a_path <- lm(DisEngmt ~ GRMS * GRIcntly, data = Lewis_df)
summ(Mod_a_path, digits = 3)
  
```

Looking at these results we can see that the predictors account for about 17% of variance in disengagement coping. However, there is no significance in the predictors. Neither the IV variable (GRMS, [X]), nor the moderator (GRIcntly, [Y]), nor its interaction (GRMS:GRIcntly, [XW]) are significant.

Observations	212
Dependent variable	DisEngmt
Type	OLS linear regression

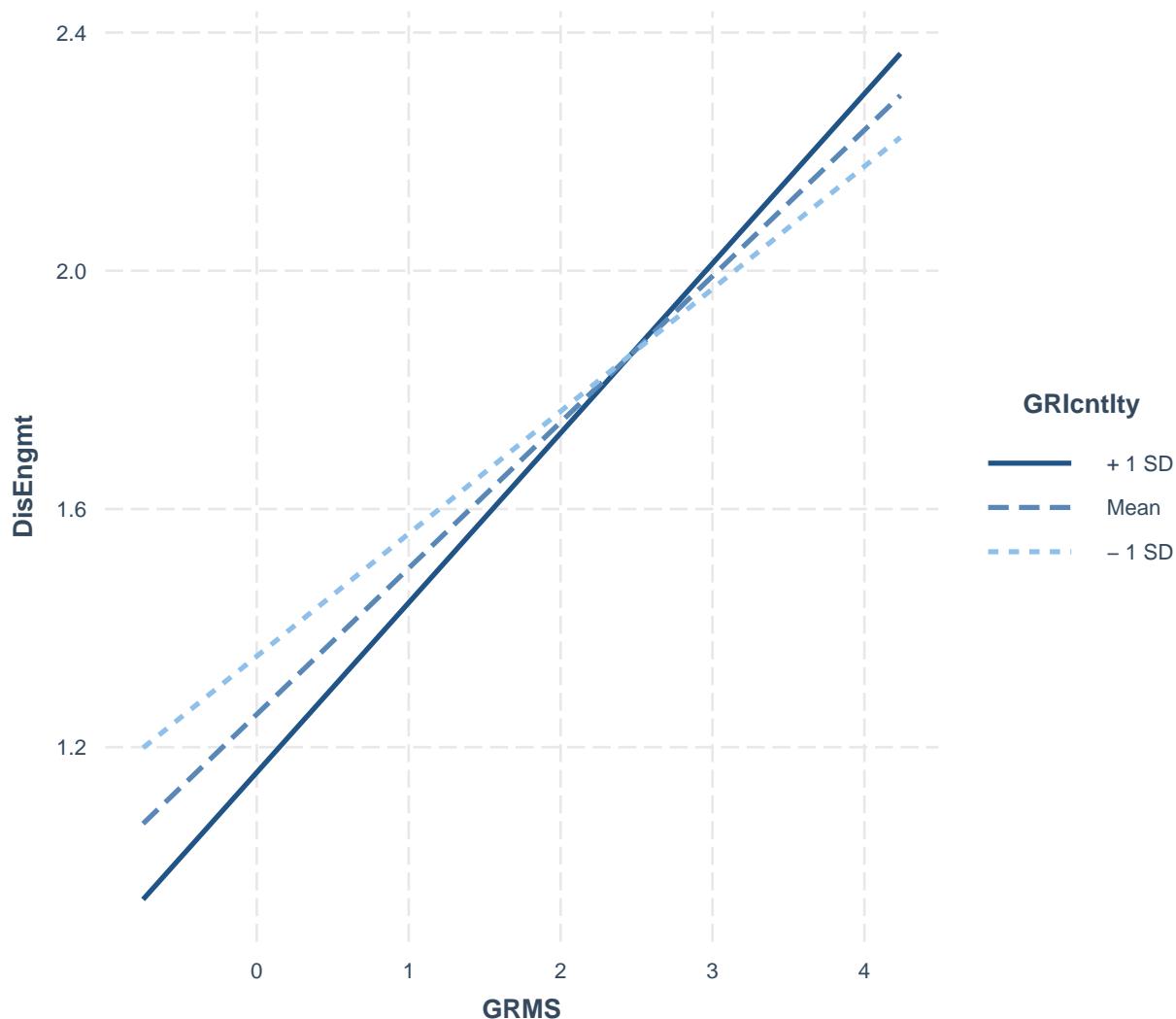
F(3,208)	14.685
R <sup>2</sup>	0.175
Adj. R <sup>2</sup>	0.163

	Est.	S.E.	t val.	p
(Intercept)	1.796	0.415	4.325	0.000
GRMS	0.025	0.184	0.138	0.890
GRIcntlty	-0.095	0.073	-1.290	0.199
GRMS:GRIcntlty	0.038	0.032	1.217	0.225

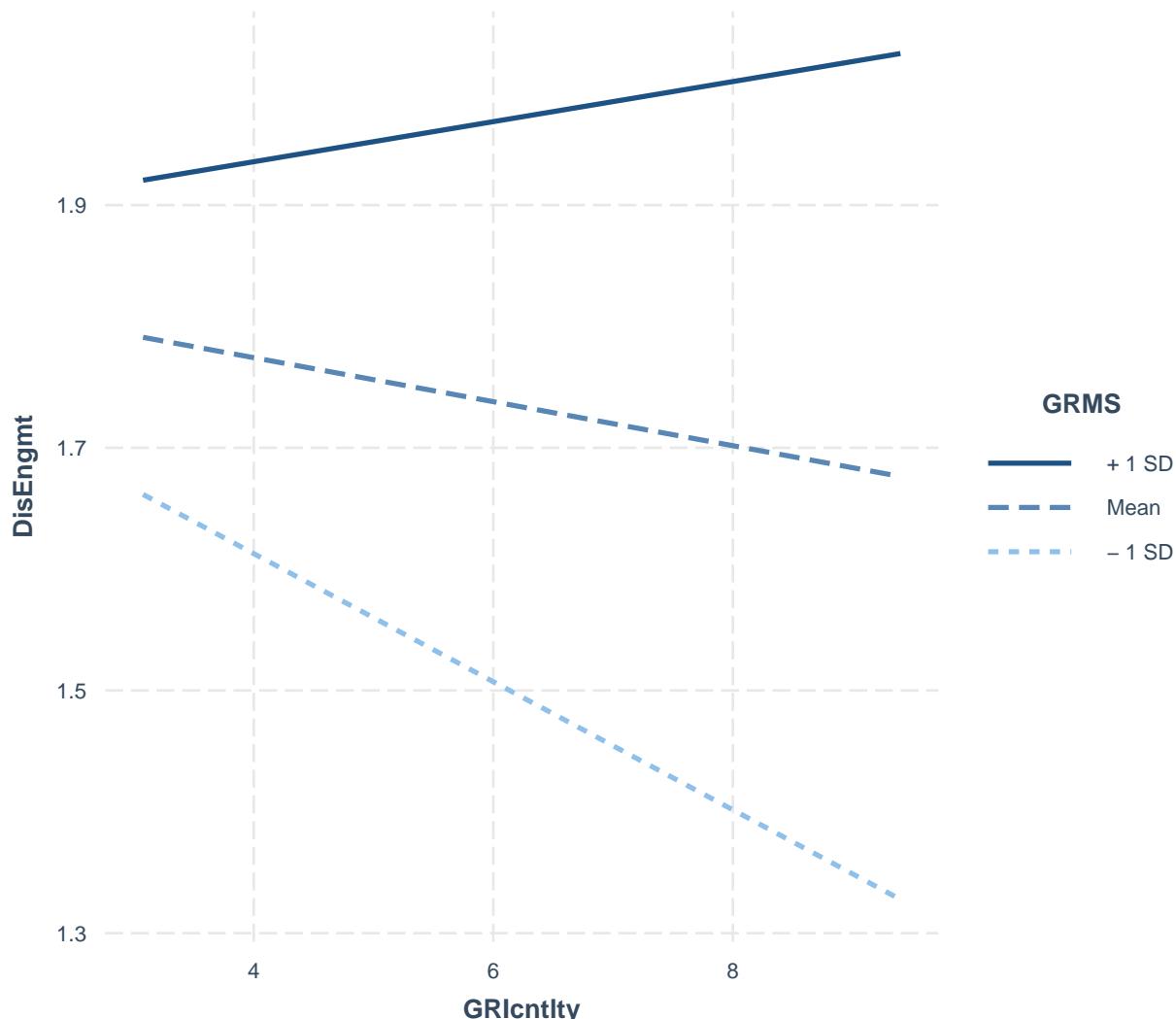
Standard errors: OLS

It's always helpful to graph the relationship. The *interaction\_plot()* function from the package, *interactions* can make helpful illustrations. In the case of interactions/moderations, I like to run them "both ways" to see which makes more sense.

```
interaction_plot(Mod_a_path, pred = GRMS, modx = GRIcntlty)
```



```
interact_plot(Mod_a_path, pred = GRIcntly, modx = GRMS)
```



The figure with GRlcntly as the moderator, shows a very similar prediction of disengagement coping from gendered racial microaggressions. The figure that uses GRMS as the moderator, visually, looks like there are big differences as a function of GRMS. Looking at the Y axis (disengagement), though, shows that the scaling variables are not well-spaced.

Next, let's probe the interaction with simple slopes. Probing the interaction is a common follow-up. With these additional inferential tests we can see where in the distribution of the moderator, X has an effect on Y that is different from zero (and where it does not). There are two common approaches.

The Johnson-Neyman is a *floodlight* approach and provides an indication of the places in the distribution of W (moderator) that X has an effect on Y that is different than zero. The pick-a-point is sometimes called the *analysis of simple slopes* or a *spotlight* approach, probes the distribution at specific values (often the  $M \pm 1SD$ ).

```
sim_slopes(Mod_a_path, pred = GRMS, modx = GRlcntly)
```

```
## JOHNSON-NEYMAN INTERVAL
##
```

```

## When GRIcntlty is INSIDE the interval [3.45, 15.76], the slope of GRMS is p
## < .05.
##
## Note: The range of observed values of GRIcntlty is [3.08, 9.40]
##
## SIMPLE SLOPES ANALYSIS
##
## Slope of GRMS when GRIcntlty = 4.68 (- 1 SD):
##
##   Est.   S.E.   t val.      p
##   ----- -----
##   0.21   0.05     4.15   0.00
##
## Slope of GRMS when GRIcntlty = 5.71 (Mean):
##
##   Est.   S.E.   t val.      p
##   ----- -----
##   0.25   0.04     6.49   0.00
##
## Slope of GRMS when GRIcntlty = 6.74 (+ 1 SD):
##
##   Est.   S.E.   t val.      p
##   ----- -----
##   0.28   0.05     5.68   0.00

# sim_slopes(Mod_a_path, pred=GRIcntlty, modx = GRMS) #sometimes I
# like to look at it in reverse -- like in the plots

```

The Johnson-Neyman suggests that between the GRIcntlty values of 3.02 and 9.03, the relationship between GRMS is statistically significant. We see the same result in the pick-a-point approach where at the GRIcntlty values of 4.68, 5.71, and 6.74, X has a statistically significant effect on Y. Is this a contradiction to the non-significant interaction effect?

No. The test of interaction is an interaction about the relationship between  $W$  and  $X$ 's effect on  $Y$ . Just showing that  $X$  is significantly related to  $Y$  for a specific value does not address any dependence upon the moderator ( $W$ ). Hayes [2018] covers this well in his Chapter 14, in the section “Reporting a Moderation Analysis.”

### What have we learned in this simple moderation?

- While there are no significant predictors (neither  $X$ ,  $W$ , nor  $XW$ ), the model accounts for about 17% of variance in the DV.
- Although there was a non-significant effect of GRMS on disengagement coping, analysis of simple slopes suggested a significant relationship between these variables at a given ranges of GRIcntlty.
- We'll keep these in mind.

### 8.5.1.2 Analysis #2: Another simple moderation

We are asking, “Does gendered racial identity centrality moderate the relationship between gendered racial microaggressions and mental health?”

Y = mental health X = gendered racial microaggressions W = GRI centrality

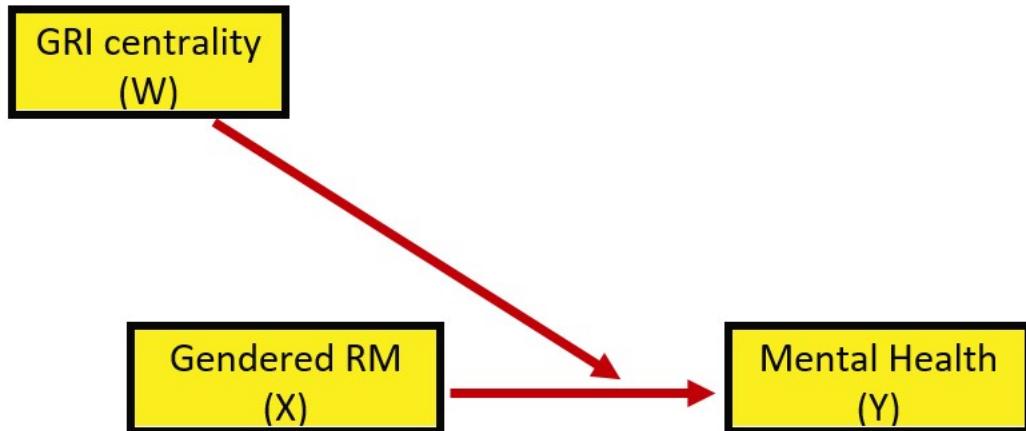


Figure 8.6: Image of statistical representation of the simple moderation estimating MntlHlth from GRMS, moderated by GRIcntlty

As before, this is our formulaic rendering:

$$Y = b_0 + b_1 X + b_2 W + b_3 XW + e_Y$$

```
Mod_c_path <- lm(MntlHlth ~ GRMS * GRIcntlty, data = Lewis_df)
summ(Mod_c_path, digits = 3)
```

Observations	212
Dependent variable	MntlHlth
Type	OLS linear regression

F(3,208)	8.050
R <sup>2</sup>	0.104
Adj. R <sup>2</sup>	0.091

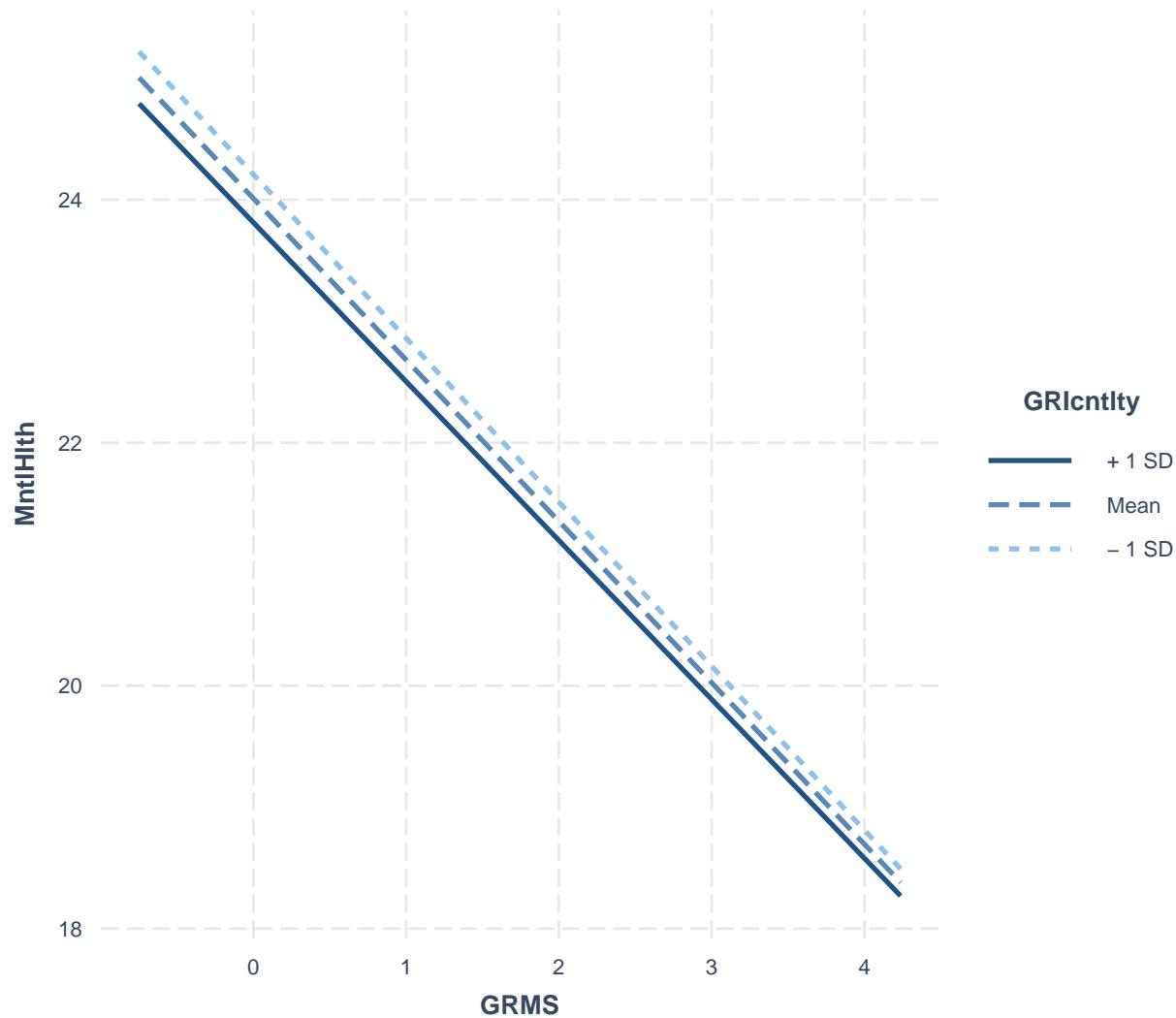
	Est.	S.E.	t val.	p
(Intercept)	25.110	3.127	8.030	0.000
GRMS	-1.442	1.386	-1.041	0.299
GRIcntlty	-0.193	0.553	-0.348	0.728
GRMS:GRIcntlty	0.020	0.238	0.084	0.933

Standard errors: OLS

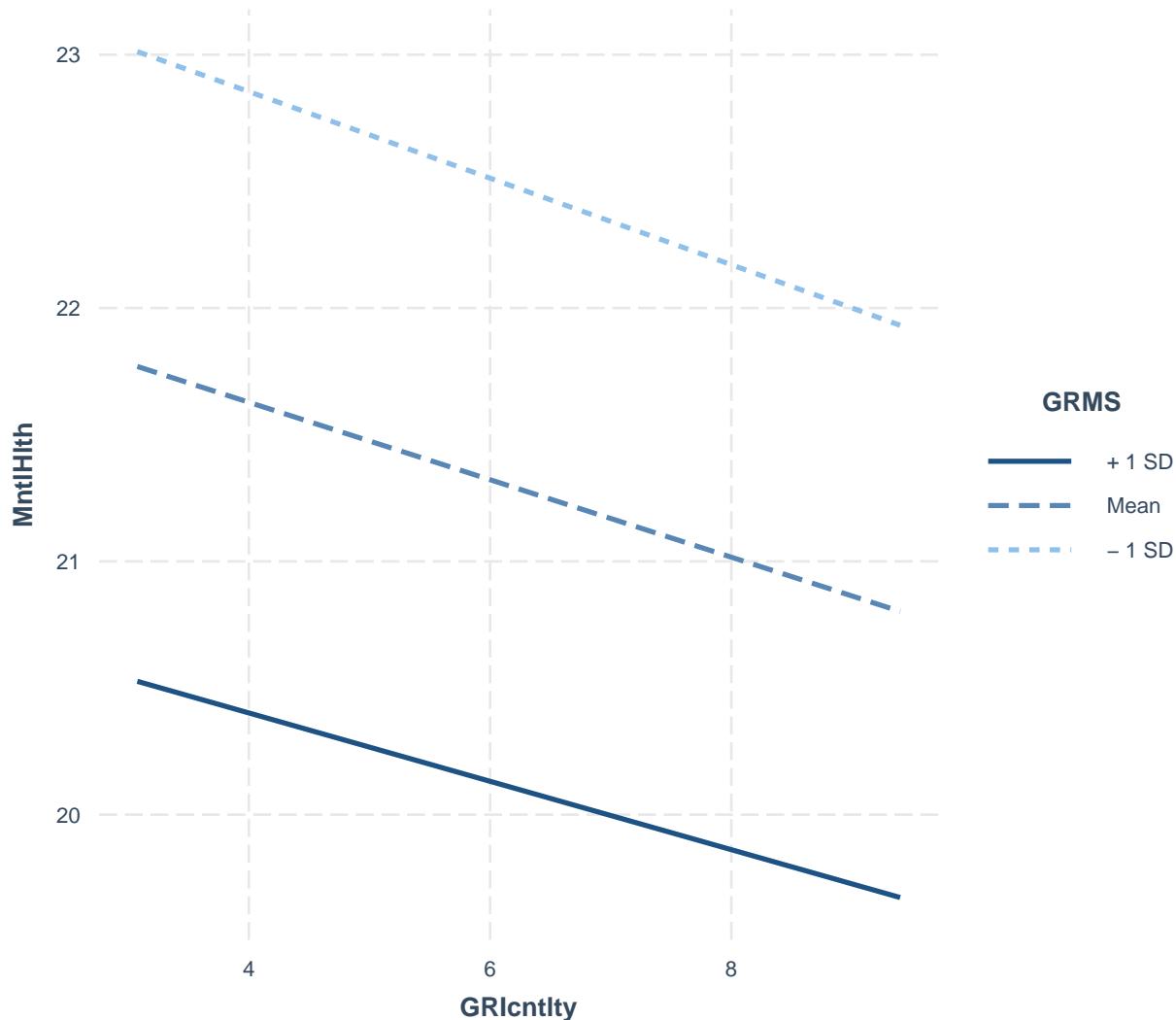
In this model that is, overall, statistically significant, we account for about 11% of variance in the DV. Looking at these results we can see that there is no significance in the predictors. Neither the IV (GRMS, [X]), nor the moderator (GRIcntlty, [Y])), nor its interaction (GRMS:GRIcntlty, [XW]) are significant.

Let's look at the plots.

```
interact_plot(Mod_c_path, pred = GRMS, modx = GRIcntlty)
```



```
interact_plot(Mod_c_path, pred = GRIcntlty, modx = GRMS)
```



The figure with GRIcntlty as the moderator, shows fanning out when mental health is high and GRMS is low.

Next, let's probe the interaction with simple slopes. Probing the interaction is a common follow-up. With these additional inferential tests we can see where in the distribution of the moderator, X has an effect on Y that is different from zero (and where it does not). There are two common approaches.

The Johnson-Neyman is a *floodlight* approach and provides an indication of the places in the distribution of W (moderator) that X has an effect on Y that is different than zero. The pick-a-point is sometimes called the *analysis of simple slopes* or a *spotlight* approach, probes the distribution at specific values (often the  $M \pm 1SD$ ).

```
sim_slopes(Mod_c_path, pred = GRMS, modx = GRIcntlty)
```

```
## JOHNSON-NEYMAN INTERVAL
##
## When GRIcntlty is INSIDE the interval [3.00, 8.15], the slope of GRMS is p
## < .05.
```

```

## 
## Note: The range of observed values of GRIcntlty is [3.08, 9.40]
## 
## SIMPLE SLOPES ANALYSIS
## 
## Slope of GRMS when GRIcntlty = 4.68 (- 1 SD):
## 
##   Est.    S.E.    t val.      p
##   ----- -----
##   -1.35   0.37    -3.61    0.00
## 
## Slope of GRMS when GRIcntlty = 5.71 (Mean):
## 
##   Est.    S.E.    t val.      p
##   ----- -----
##   -1.33   0.28    -4.67    0.00
## 
## Slope of GRMS when GRIcntlty = 6.74 (+ 1 SD):
## 
##   Est.    S.E.    t val.      p
##   ----- -----
##   -1.31   0.38    -3.46    0.00

```

```

# sim_slopes(Mod_c_path, pred=GRIcntlty, modx = GRMS) #sometimes I
# like to look at it in reverse -- like in the plots

```

The Johnson-Neyman suggests that between the GRIcntlty values of 2.972 and 7.46, the relationship between GRMS is statistically significant. We see the same result in the pick-a-point approach where at the GRIcntlty values of 4.68, 5.71, and 6.74, X has a statistically significant effect on Y. Is this a contradiction to the non-significant interaction effect?

Again. No. The test of interaction is an interaction about the relationship between  $W$  and  $X$ 's effect on  $Y$ . Just showing that  $X$  is significantly related to  $Y$  for a specific value does not address any dependence upon the moderator ( $W$ ). Hayes [2018] covers this well in his Chapter 14, in the section “Reporting a Moderation Analysis.”

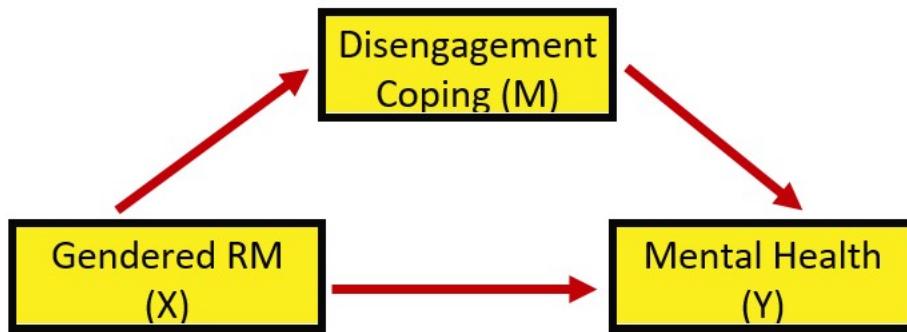
### What have we learned in this simple moderation?

- As predictors to the DV, disengagement coping, the IV (X), moderator (W), and its interaction term (XW) have non-significant effects. That said, the overall model was significant and accounted for 11% of variance in the DV.
- Although there was a non-significant effect of GRMS on mental health, analysis of simple slopes suggested a significant relationship between these variables at a given range of GRIcntlty.
- We'll keep these in mind.

### 8.5.1.3 Analysis #3: A simple mediation

We are asking, “Does disengagement coping mediate the relationship between gendered racial microaggressions and mental health?”

$Y$  = mental health  $X$  = gendered racial microaggressions  $M$  = GRI centrality



Looking at the diagram,

with two consequent variables (i.e., those with arrows pointing to them) we can see two equations are needed to explain the model:

$$M = i_M + aX + e_M$$

$$Y = i_Y + c'X + bM + e_Y$$

To conduct this analysis, I am using the guidelines in the [chapter on simple mediation](#). We are switching to the *lavaan* package.

```

library(lavaan)
set.seed(210421) #reset in case you choose to separate these sections
LMedModel <- "
  MntlHlth ~ b*DisEngmt + c_p*GRMS
  DisEngmt ~ a*GRMS

  #intercepts
  DisEngmt ~ DisEngmt.mean*1
  MntlHlth ~ MntlHlth.mean*1

  indirect := a*b
  direct := c_p
  total_c := c_p + (a*b)
  "

LMed_fit <- sem(LMedModel, data = Lewis_df, se = "bootstrap", missing = "fiml")
LMed_Sum <- summary(LMed_fit, standardized = T, rsq = T, ci = TRUE)
LMed_ParEsts <- parameterEstimates(LMed_fit, boot.ci.type = "bca.simple",
  standardized = TRUE)
LMed_Sum
  
```

```

## lavaan 0.6.16 ended normally after 1 iteration
##
##   Estimator                      ML
## Optimization method            NLMINB
## Number of model parameters      7
##
##   Number of observations        212
## Number of missing patterns       1
##
## Model Test User Model:
##
##   Test statistic                 0.000
## Degrees of freedom                  0
##
## Parameter Estimates:
##
##   Standard errors                Bootstrap
## Number of requested bootstrap draws    1000
## Number of successful bootstrap draws   1000
##
## Regressions:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## MntlHlth ~
##   DisEngmt (b)   -3.551   0.438  -8.101  0.000  -4.395  -2.649
##   GRMS   (c_p)  -0.504   0.238  -2.123  0.034  -0.981  -0.043
## DisEngmt ~
##   GRMS     (a)   0.241   0.035   6.884  0.000   0.170   0.305
## Std.lv  Std.all
##
##   -3.551  -0.491
##   -0.504  -0.119
##
##   0.241   0.410
##
## Intercepts:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .DsEngmt (DsE.)  1.270   0.078  16.330  0.000   1.117   1.432
## .MntlHlt (MnH.) 28.588   0.692  41.330  0.000  27.231  29.910
## Std.lv  Std.all
##   1.270   2.401
##   28.588   7.482
##
## Variances:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
## .MntlHlth      10.172   0.843  12.059  0.000   8.375  11.723
## .DisEngmt       0.233   0.021  10.883  0.000   0.189   0.274
## Std.lv  Std.all
##   10.172   0.697

```

```

##      0.233    0.832
##
## R-Square:
##              Estimate
##   MntlHlth       0.303
##   DisEngmt       0.168
##
## Defined Parameters:
##             Estimate Std.Err z-value P(>|z|) ci.lower ci.upper
##   indirect     -0.857  0.169 -5.073  0.000 -1.217  -0.544
##   direct       -0.504  0.238 -2.122  0.034 -0.981  -0.043
##   total_c      -1.362  0.250 -5.443  0.000 -1.830  -0.858
##   Std.lv  Std.all
##   -0.857  -0.201
##   -0.504  -0.119
##   -1.362  -0.320

```

**LMed\_ParEsts**

##	lhs	op	rhs	label	est	se	z	pvalue	ci.lower	ci.upper
## 1	MntlHlth	~	DisEngmt	b	-3.551	0.438	-8.101	0.000	-4.466	
## 2	MntlHlth	~	GRMS	c_p	-0.504	0.238	-2.123	0.034	-1.044	
## 3	DisEngmt	~	GRMS	a	0.241	0.035	6.884	0.000	0.176	
## 4	DisEngmt	~1		DisEngmt.mean	1.270	0.078	16.330	0.000	1.113	
## 5	MntlHlth	~1		MntlHlth.mean	28.588	0.692	41.330	0.000	27.255	
## 6	MntlHlth	~~	MntlHlth		10.172	0.843	12.059	0.000	8.730	
## 7	DisEngmt	~~	DisEngmt		0.233	0.021	10.883	0.000	0.193	
## 8	GRMS	~~	GRMS		0.806	0.000	NA	NA	0.806	
## 9	GRMS	~1			1.990	0.000	NA	NA	1.990	
## 10	indirect	:=	a*b	indirect	-0.857	0.169	-5.073	0.000	-1.247	
## 11	direct	:=	c_p	direct	-0.504	0.238	-2.122	0.034	-1.044	
## 12	total_c	:=	c_p+(a*b)	total_c	-1.362	0.250	-5.443	0.000	-1.865	
##	ci.upper	std.lv	std.all	std.nox						
## 1	-2.687	-3.551	-0.491	-0.491						
## 2	-0.076	-0.504	-0.119	-0.132						
## 3	0.313	0.241	0.410	0.457						
## 4	1.427	1.270	2.401	2.401						
## 5	29.923	28.588	7.482	7.482						
## 6	12.180	10.172	0.697	0.697						
## 7	0.277	0.233	0.832	0.832						
## 8	0.806	0.806	1.000	0.806						
## 9	1.990	1.990	2.216	1.990						
## 10	-0.568	-0.857	-0.201	-0.224						
## 11	-0.076	-0.504	-0.119	-0.132						
## 12	-0.891	-1.362	-0.320	-0.356						

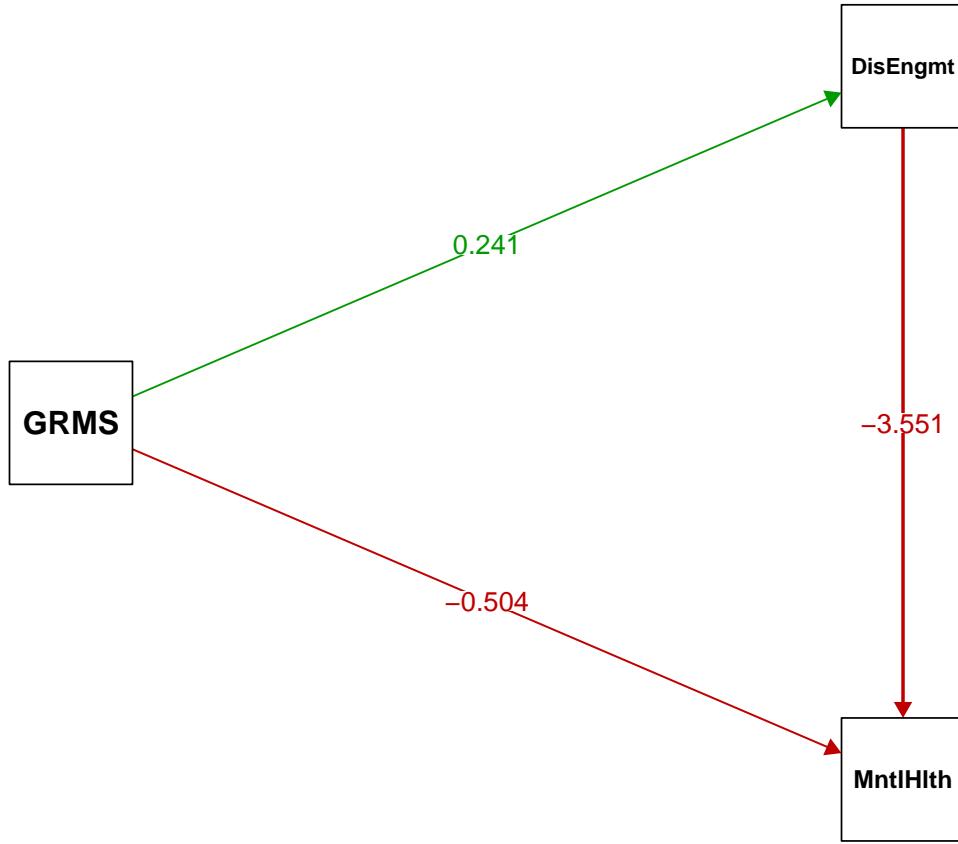
In this simple mediation we learn\*:

- The  $a$  path (GRMS → DisEngmt) is statistically significant.
- The  $b$  path (DisEngmt → MntlHlth) is statistically significant.
- The total effect (GRMS → MntlHlth) is statistically significant.
- The direct effect (GRMS → MntlHlth when DisEngmt is in the model) falls out of significance.
- The indirect effect is statistically significant.
- The model accounts for 30% of the variance in mental health (DV) and 17% of the variance in disengagement coping (M).

Recall how the bootstrapped, bias-corrected confidence intervals can be different? It's always good to check. In this case, CI95s and the  $p$  values are congruent.

```
set.seed(210421)
library(semPlot)
semPaths(LMed_fit, #must identiy the model you want to map
         what = "est", ##"est" plots the estimates, but keeps it greyscale with no fading
#whatLabels = "stand", ##"stand" changes to standardized values
         layout = 'tree', rotation = 2, #together, puts predictors on left, IVs on right
#layout = 'circle',
         edge.label.cex = 1.00, #font size of parameter values
#edge.color = "black", #overwrites the green/black coloring
         sizeMan=10, #size of squares/observed/"manifest" variables
         fade=FALSE, #if TRUE, there lines are faded such that weaker lines correspond with low
         esize=2,
         asize=3,
         #label.prop = .5,
         label.font = 2.5, #controls size (I think) of font for labels
         label.scale = TRUE, #if false, the labels will not scale to fit inside the nodes
         nDigits = 3, #decimal places (default is 2)
         residuals = FALSE,#excludes residuals (and variances) from the path diagram
         nCharNodes = 0, #specifies how many characters to abbreviate variable lables; default
         intercepts = FALSE, #gets rid of those annoying triangles (intercepts) in the path diagram
)
title("Disengagement Coping as Mediator between GRMS and Mental Health")
```

### Disengagement Coping as Mediator between GRMS and Mental Health



## 8.6 The Moderated Mediation: A Combined analysis

For a quick reminder, the diagram with labeled paths will help specify this in *lavaan*.

Looking at the diagram, with two consequent variables (i.e., those with arrows pointing to them) we can see two equations are needed to explain the model:

$$M = i_M + a_1 X + a_2 W + a_3 XW + e_M$$

$$Y = i_Y + c'_1 X + c'_2 W + c'_3 XW + bM + e_Y$$

Y = MntlHlth X = GRMS W = DisEngmt M = GRIcntly

### 8.6.1 Specification in *lavaan*

In the code below

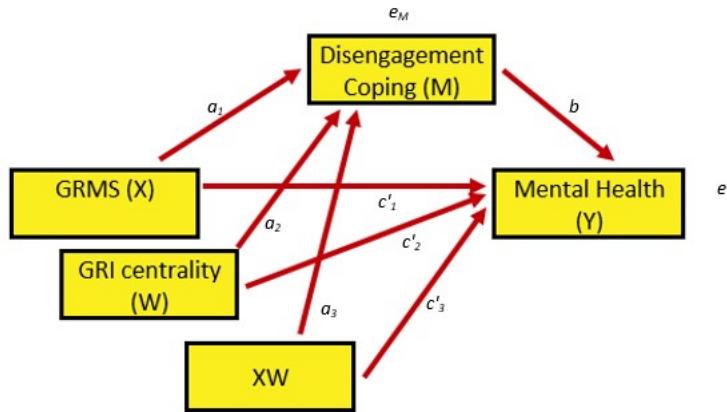


Figure 8.7: Image of statistical representation of the conditional process analysis model where the moderator is hypothesized to change the a and c' paths

- First specify the equations, hints
  - the a,b,c, labels are affixed with the \*(asterisk)
  - interaction terms are identified with the colon
- Create code for the intercepts (Y and M) with the form: VarName ~ VarName.mean\*1
- Create code for the mean and variance of all moderators (W, Z, etc.); these will be used in simple slopes.
  - Means use the form: VarName ~ VarName.mean\*1
  - Variances use the form: VarName ~~VarName.var\*VarName
- Calculate the *index of moderated mediation*: quantifies the relationship between the moderator and the indirect effect.
  - To the degree that the value of the IMM is different from zero and the associated inferential test is statistically significant (bootstrapped confidence intervals are preferred; more powerful), we can conclude that the indirect effect is moderated.
    - \* The IMM is used in the formula to calculate the conditional indirect effects.
    - \* Hayes argues that a statistically significant IMM suggest they are (boom, done, p. 430).
- Create code to calculate indirect effects conditional on ( $M +/- 1SD$ ) moderator with the general form:
  - product of the indirect effect ( $a*b$ ) PLUS
  - the product of the IMM and the moderated value
- Because our direct path is moderated, we will use a similar process to specify the direct effects conditional on ( $M +/- 1SD$ ) moderator with the general form:
  - the direct effect ( $c_p1$ ) PLUS
  - the moderated value ( $c_p3$ ) at each of the three levels ( $M +/- 1SD$ )

- Although they don't tend to be reported, you can create total effects conditional on the ( $M \pm 1SD$ ). These are simply the sum of the  $c_p$  and all indirect paths, specified individually, at their  $M \pm 1SD$  conditional values.

```

set.seed(190505)
Combined <- '
  #equations
  DisEngmt ~ a1*GRMS + a2*GRICntlty + a3*GRMS:GRICntlty
  MntlHlth ~ c_p1*GRMS + c_p2*GRICntlty + c_p3*GRMS:GRICntlty + b*DisEngmt

  #intercepts
  DisEngmt ~ DisEngmt.mean*1
  MntlHlth ~ MntlHlth.mean*1

  #means, variances of W for simple slopes
  GRICntlty ~ GRICntlty.mean*1
  GRICntlty ~~ GRICntlty.var*GRICntlty

  #index of moderated mediation, there will be an a and b path in the product
  #if the a and/or b path is moderated, select the label that represents the moderation
  imm := a3*b

  #Note that we first create the indirect product, then add to it the product of the imm and
  indirect.SDbelow := a1*b + imm*(GRICntlty.mean - sqrt(GRICntlty.var))
  indirect.mean := a1*b + imm*(GRICntlty.mean)
  indirect.SDabove := a1*b + imm*(GRICntlty.mean + sqrt(GRICntlty.var))

  #direct effect is also moderated so calculate with c_p1 + c_p3
  direct.SDbelow := c_p1 + c_p3*(GRICntlty.mean - sqrt(GRICntlty.var))
  direct.Smean := c_p1 + c_p3*(GRICntlty.mean)
  direct.SDabove := c_p1 + c_p3*(GRICntlty.mean + sqrt(GRICntlty.var))

  #total effect
  total.SDbelow := direct.SDbelow + indirect.SDbelow
  total.mean := direct.Smean + indirect.mean
  total.SDabove := direct.SDabove + indirect.SDabove
'

Combined_fit <- sem(Combined, data = Lewis_df, se = "bootstrap", missing = 'fiml', bootstrap = 1000)

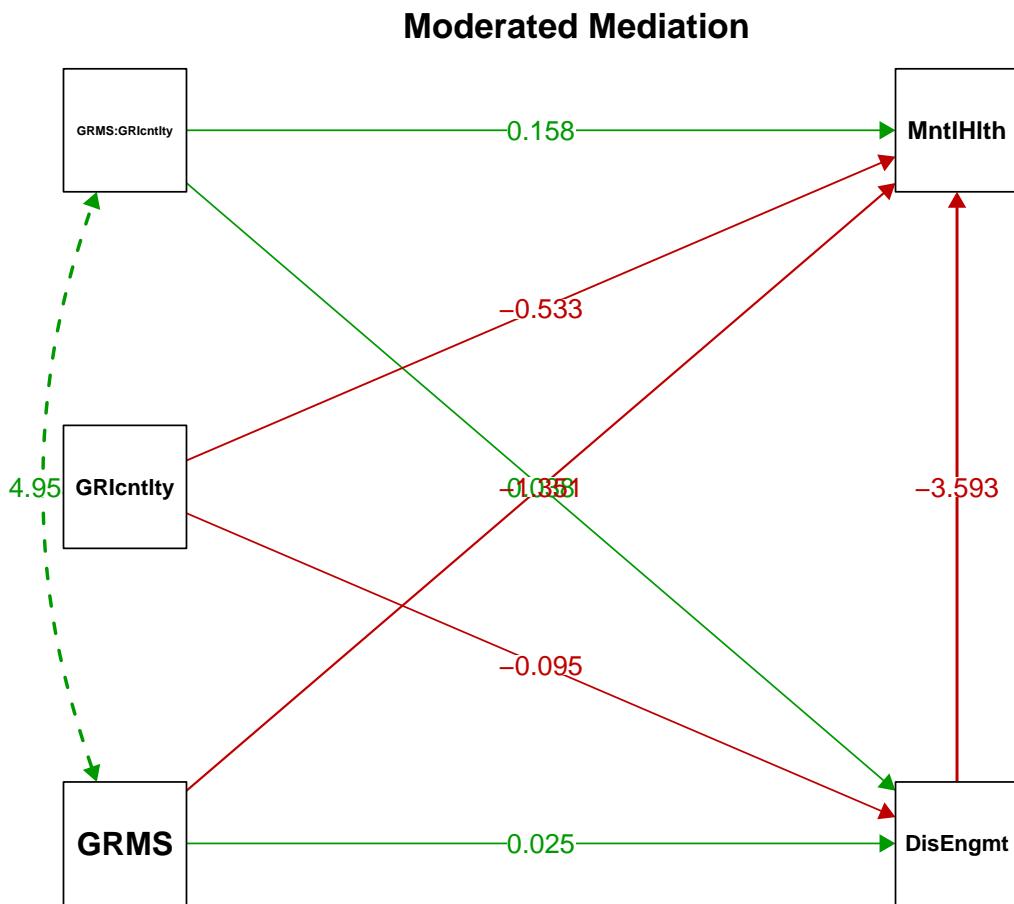
## Warning in lav_partable_vnames(FLAT, "ov.x", warn = TRUE): lavaan WARNING:
##   model syntax contains variance/covariance/intercept formulas
##   involving (an) exogenous variable(s): [GRICntlty]; These variables
##   will now be treated as random introducing additional free
##   parameters. If you wish to treat those variables as fixed, remove
##   these formulas from the model syntax. Otherwise, consider adding
##   the fixed.x = FALSE option.

```

```
cFITsum <- summary(Combined_fit, standardized = TRUE, rsq=T, ci=TRUE)
cParamEsts <- parameterEstimates(Combined_fit, boot.ci.type = "bca.simple", standardized=TRUE)
```

### 8.6.2 A quick plot

```
library(semPlot)
semPaths(Combined_fit, #must identify the model you want to map
         what = "est", ##"est" plots the estimates, but keeps it greyscale with no fading
#whatLabels = "stand", ##"stand" changes to standardized values
         layout = 'tree', rotation = 2, #together, puts predictors on left, IVs on right
#layout = 'circle',
         edge.label.cex = 1.00, #font size of parameter values
#edge.color = "black", #overwrites the green/black coloring
         sizeMan=10, #size of squares/observed/"manifest" variables
         fade=FALSE, #if TRUE, there lines are faded such that weaker lines correspond with lower p-values
         esize=2,
         asize=3,
         #label.prop = .5,
         label.font = 2.5, #controls size (I think) of font for labels
         label.scale = TRUE, #if false, the labels will not scale to fit inside the nodes
         nDigits = 3, #decimal places (default is 2)
         residuals = FALSE,#excludes residuals (and variances) from the path diagram
         nCharNodes = 0, #specifies how many characters to abbreviate variable labels; default is 10
         intercepts = FALSE, #gets rid of those annoying triangles (intercepts) in the path diagram
)
title("Moderated Mediation")
```



```

write.csv(cParamEsts, file = "Combined_fit.csv") #optional to write it to a .csv file

##
## Attaching package: 'formattable'

## The following object is masked from 'package:MASS':
##
##     area
  
```

### 8.6.3 Beginning the interpretation

We have already looked at some of the simple effects found within the more complex model. Let's grab the formulae.

$$\hat{M} = i_M + a_1 X + a_2 W + a_3 XW + e_M$$

$$\hat{Y} = i_Y + c'_1 X + c'_2 W + c'_3 XW + bM + e_Y$$

And substitute in our values

$$\begin{aligned}\hat{M} &= 1.417 + 0.212X + (-0.027)W + 0.006XW \\ \hat{Y} &= 31.703 + (-1.4115)X + (-0.556)W + 0.164XW + (-3.567)M\end{aligned}$$

#### 8.6.4 Tabling the data

Table 1

---

Analysis of Moderated Mediation for GRMS, Gendered Racial Identity Centrality, Coping, and Mental Health

---

	Disengagement Coping (M)				Mental Health (Y)			
Antecedent	path	B	SE	p	path	B	SE	p
constant	$i_M$	1.796	0.406	0.000	$i_Y$	31.564	2.483	0.000
GRMS (X)	$a_1$	0.025	0.174	0.884	$c_1$	-1.351	0.994	0.174
GRICntrlty (W)	$a_2$	-0.095	0.073	0.196	$c_2$	-0.533	0.419	0.203
GRMS*GRICntrlty(XW)	0.038	0.031	0.208		$c_3$	0.158	0.168	0.347
DisEngmt (M)					$b$	-3.593	0.424	0.000

---

$$R^2 =$$


---

---

Conditional Indirect, Direct, and Total Effects at Gendered Racial Identity Centrality Values

---

	Boot effect	Boot SE	Boot CI95 lower	Boot CI95 upper
Index of moderated mediation	-0.138	0.110	-0.360	0.074
Indirect				
-1 SD	-0.739	0.184	-1.161	-0.412
Mean	-0.881	0.165	-1.281	-0.606
+1 SD	-1.023	0.212	-1.467	-0.632
Direct				
-1 SD	-0.610	0.298	-1.211	-0.026

Mean	-0.447	0.241	-0.896	0.034
+1 SD	-0.285	0.292	-0.852	0.323
Total				
-1 SD	-1.349	0.310	-1.972	-0.785
Mean	-1.329	0.255	-1.846	-0.839
+1 SD	-1.308	0.338	-1.922	-0.591

*Note.* IV(X) = gendered racial microaggressions; M = disengagement coping; W; gendered racial identity centrality; Y = mental health. The significance of the indirect and direct effects were calculated with bias-corrected confidence intervals (.95) bootstrap analysis.

of the variance in disengagement coping (mediator) and of the variance in mental health (DV) are predicted by their respective models.

The model we tested suggested that both the indirect and direct effects should be moderated. Hayes provides a more detailed and elaborate explanation of this on pp. 447 - 458.

#### 8.6.4.1 Conditional Indirect effects

- An indirect effect can be moderated if either the *a* or *b* path (or both) is(are) moderated.
- If at least one of the indirect paths is part of a moderation, then the whole indirect (axb) path would be moderated.
  - In this model, we specified a moderation of the *a* path.
- We know if the *a* path is moderated if the moderation term is statistically significant.
  - In our case,  $a_3$  GRMS:GRIcntlty was not statistically significant ( $B = 0.038, p = 0.208$ ).
- We also look at the *Index of Moderated Mediation*. The IMM is the product of the moderated path (in this case, the value of  $a_3$ ) and *b*. If this index is 0, then the slope of the line for the indirect effect is flat. The bootstrap confidence interval associated with this test is the way to determine whether/not this slope is statistically significant from zero. In our case, IMM = -0.138, CI095 = -0.360 to 0.074. This suggests that we do not have a moderated mediation. Hayes claims the IMM saves us from formally comparing (think “contrasts” pairs of conditional indirect effects)
- We can even get more information about the potentially moderated indirect effect by *probing the conditional indirect effect*. Because an indirect effect is not normally distributed, Hayes discourages using a Johnson-Neyman approach and suggests that we use the pick-a-point. He usually selects the 16th, 50th, and 84th percentiles of the distribution. However, many researchers commonly report the mean+/-1SD.
  - at 1SD below the mean  $B = -0.739$ , CI95 -1.161 to -0.412;
  - at the mean  $B = -0.881$ , CI95 -1.281 to -0.606).
- at 1SD above the mean, the conditional indirect effect was significant ( $B = -1.023$ , CI95 -1.467 to -0.632).

- Looking at the relative consistency of the  $B$  weights and the consistently significant  $p$  values, we see that there was an indirect effect throughout the varying levels of the moderator, gendered racial identity centrality. Thus, it makes sense that this was not a moderated mediation.

#### 8.6.4.2 Conditional Direct effect

- The direct effect of X to Y estimates how differences in X relate to differences in Y holding constant the proposed mediator(s).
- We know the direct effect is moderated if the interaction term ( $c_{\text{p3}}$ ) is statistically significant. In our case, it was not ( $B = 0.158$ ,  $p = 0.347$ ).
- Probing a conditional direct effect is straightforward...we typically use the same points as we did in the probing of the conditional indirect effect.
  - For both my values (mean and  $\pm 1\text{SD}$ ) and Hayes values (16th, 50th, 84th percentiles), the direct effect (e.g., the effect of skepticism on willingness to donate) was not statistically significant from zero at any level of the moderator.

#### 8.6.5 Model trimming

Hayes terms it *pruning*, but suggests that when there is no moderation of an effect, the researcher may want to delete that interaction term. In our case, neither the direct nor indirect effect was moderated (although the  $+1\text{SD}$  was close ( $B = -0.285$ ,  $p = 0.323$ )). Deleting these paths one at a time is typical practice because the small boost of power with each deleted path may “turn on” significance elsewhere. If I were to engage in model trimming, I would start with the indirect effect to see if the direct effect became moderated. This is consistent with the simple moderation we ran earlier where we saw a fanning out at one end of the distribution.

#### 8.6.6 APA Style Write-up

*Note:* Make sure to look at the write-up in the Lewis et al. [Lewis et al., 2017] manuscript. I am a little confused in that Figure 2 of their manuscript suggests there was a moderation of both the  $a$  and  $c'$  paths. However, the results in Table 4 do not provide information about the moderation to the  $c'$  path. The Lewis et al. write-up is an efficient one, simultaneously presenting the results of two outcome variables – mental and physical health. While our  $B$  weights from our simulated data map similarly onto those reported in the Lewis et al. manuscript, we do not get the statistically significant moderated mediation that they get.

### Results

#### Preliminary Analyses

- Missing data analysis and managing missing data
- Bivariate correlations, means, SDs
- Distributional characteristics, assumptions, etc.
- Address limitations and concerns

**Primary Analyses** Our analysis evaluated a moderation mediation model predicting mental health ( $Y/\text{MntlHlth}$ ) from gendered racial microaggressions ( $X/\text{GRMS}$ ) mediated by disengagement coping ( $M/\text{DisEngmt}$ ). Gendered racial identity centrality ( $W/\text{GRICntrlty}$ ) was our moderating variable. We specified a moderation of path  $a$  ( $X/\text{GRMS}$  to  $M/\text{DisEngmt}$ ) and the direct path,  $c'$  ( $X/\text{GRMS}$  to  $Y/\text{MntlHlth}$ ). Data were analyzed with maximum likelihood estimation in the R package *lavaan* (v. 0.6-7); the significance of effects were tested with 1000 bootstrap confidence intervals. Results of the full model are presented in Table 1 and illustrated in Figure 1 (*a variation of the semPlot or Hayes style representation*). The formula for the mediator and dependent variable are expressed below.

$$\hat{M} = 1.417 + 0.212X + (-0.027)W + 0.006XW$$

$$\hat{Y} = 31.703 + (-1.4115)X + (-0.556)W + 0.164XW + (-3.567)M$$

Results suggested a strong and negative total effect of gendered racial microaggressions on mental health that is mediated through disengagement coping. That is, in the presence of gendered racial microaggressions, participants increased disengagement coping which, in turn, had negative effects on mental health. The index of moderated mediation was -0.138 (CI95 -0.360 to 0.074) and suggested that the indirect effects were not conditional on the values of the moderator. While there was no evidence of moderation on the indirect or direct paths, there was a statistically significant, and consistently strong, mediation throughout the range of the gendered racial identity centrality (moderator). \*\*Because we did not have a moderated mediation, I would probably not include the rest of this paragraph (nor include the moderation figure). I just wanted to demonstrate how to talk about findings if they were significant (although I acnowledg throughat that these are non-significant).\* Figure 2 illustrates the conditional effects (non-significant) of GRMS (X) on mental health (Y) among those at the traditional levels of mean and +/- 1 SD where there is a fanning out of the effect of GRMS on when the presence of gendered racial microaggressions is low and mental health is at its highest. In this combination, mental health is even more positive in the presence of positive gendered racial identity centrality. Our model accounted for of the variance in the mediator (disengagement coping) and of the variance in the dependent variable (mental health).

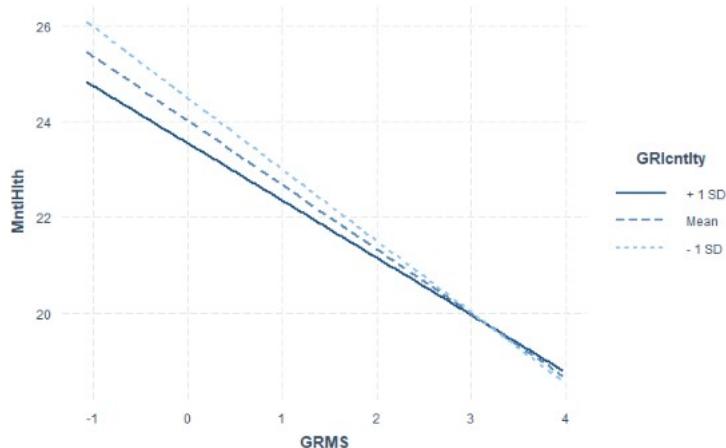


Figure 8.8: Figure 2. The non-significant moderating effect of gendered racial identity centrality on the relationship between gendered racial microaggressions and mental health

## 8.7 Residual and Related Questions...

..that you might have; or at least I had, but if had answered them earlier it would have disrupt the flow.

1. Would you stop here? Or keep tinkering?
  - I am tempted (but out of time, at least today, so stay tuned) to delete moderation of the indirect effect to see if I can get a moderated direct effect. My choice would also depend on to what I committed in any kind of pre-registration. My approach to science tends to be *model generating* [Joreskog, 1993] and in his text, Hayes [2018] advised authors to write about what they found – not all the things they tried. This *tinkering* remains strongly in the vein of theoretically driven analyses.
2. The output we get is different from the output in the journal article being used as the research vignette. Why? And should we worry about it?
  - We are simulating data. This gives us some advantages in that (unless we specify it), we never have missingness and our variables should be normally distributed. Because we are working from means, standard deviations, and correlations, our data will never be the same as the original researcher. That said, we can compare our results to the journal to *check out work*. I am somewhat reassured that our *B* weights align and somewhat concerned that the index of moderated moderation was so far off. I suppose I will always doubt myself, and will therefore be open to anyone who finds an error in the specification of the model.
3. Some of the statistics you are reporting are different than the ones in Hayes and the ones that use the PROCESS macro (e.g., what happened to the *F* test)?
  - The default estimator for *lavaan* is maximum likelihood (ML) and Hayes uses ordinary least squares (OLS). This affects both the values of coefficients, standard errors, AND the type of statistics that are reported.
  - You can ask for OLS regression by adding the statement “estimator =”GLS”. Even with this option, I have not discovered a way to obtain the *F* tests for the overall model. Researchers seem to be comfortable with this, even asking for less than we did (e.g., many do not request R square).
  - Best I can tell, researchers who do want this might use a combination of packages, using GLS estimators in *lavaan* (this easily gets them the bootstrapped CIs) and the move to a different regression package to get the intercepts and *F* tests. If I did this I would triple check to make sure that all the output really lined up.
4. Why did we ignore the traditional fit statistics associated with structural equation modeling (e.g., CFI, RMSEA).
  - I hesitate to do this with models that do not include latent variables. Therefore, we asked for an “in-between” amount of info that should be sufficient for publication submission (any editor may have their own preferences and ask for more).
5. What if I have missing data?
  - When we enter the *lavaan* world we do get options other than multiple imputation. In today’s example we used the “sem” fitting function. Unless otherwise specified, listwise

deletion (deleting the entire case when one of its variables is used to estimate the model) is the default in *lavaan*. If data are MCAR or MAR, you can add the argument *missing* = “*ml*” (or its alias *missing* = “*fiml*”). More here <https://users.ugent.be/~yrosseel/lavaan/lavaan2.pdf> on the 1.7/Missing data in lavaan slide.

- That said, the type of estimator matters. If you estimate your data with GLS (generalized least squares) or WLS (weighted least squares), you are required to have complete data (however you got it). We used maximum likelihood and, even though we had non-missing data, I used the *missing* = “*fiml*” code.

## 8.8 Practice Problems

The three problems described below were designed to grow during the series of chapters on simple and complex mediation, complex moderation, and conditional process analysis (i.e., this chapter). I have recommended that you select a dataset that includes at least four variables. If you are new to this topic, you may wish to select variables that are all continuously scaled. The IV and moderator (next chapters) could be categorical (if they are dichotomous, please use 0/1 coding; if they have more than one category it is best if they are ordered). You will likely encounter challenges that were not covered in this chapter. Search for and try out solutions, knowing that there are multiple paths through the analysis.

The suggested practice problem for this chapter is to conduct a simple mediation.

- There are a number of variables in the dataset. Swap out one or more variables in the model of simple mediation and compare your solution to the one in the chapter.
- Conduct a simple mediation with data to which you have access. This could include data you simulate on your own or from a published article.

### 8.8.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed

If this topic feels a bit overwhelming, simply change the random seed in the data simulation, then rework the problem. This should provide minor changes to the data (maybe in the second or third decimal point), but the results will likely be very similar.

---

Assignment Component		
1. Assign each variable to the X, Y, M, or W roles (ok but not required to include a cov)	5	_____
2. Specify and run the lavaan model	5	_____
3. Use semPlot to create a figure	5	_____
4. Create a table that includes regression output for the M and Y variables and the moderated effects	5	_____
5. Represent your work in an APA-style write-up	5	_____
6. Explanation to grader	5	_____
7. Be able to hand-calculate the indirect, direct, and total effects from the a, b, & c' paths	5	_____

---

Assignment Component		
<b>Totals</b>	35	_____

---

### 8.8.2 Problem #2: Rework the research vignette, but swap one or more variables

Use the simulated data, but select one of the other models that was evaluated in the Lewis et al. [Lewis et al., 2017] study. Compare your results to those reported in the manuscript.

---

Assignment Component		
1. Assign each variable to the X, Y, M, or W roles (ok but not required to include a cov)	5	_____
2. Specify and run the lavaan model	5	_____
3. Use semPlot to create a figure	5	_____
4. Create a table that includes regression output for the M and Y variables and the moderated effects	5	_____
5. Represent your work in an APA-style write-up	5	_____
6. Explanation to grader	5	_____
7. Be able to hand-calculate the indirect, direct, and total effects from the a, b, & c' paths	5	_____
<b>Totals</b>	35	_____

---

### 8.8.3 Problem #3: Use other data that is available to you

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete a simple mediation.

---

Assignment Component		
1. Assign each variable to the X, Y, M, or W roles (ok but not required to include a cov)	5	_____
2. Specify and run the lavaan model	5	_____
3. Use semPlot to create a figure	5	_____
4. Create a table that includes regression output for the M and Y variables and the moderated effects	5	_____
5. Represent your work in an APA-style write-up	5	_____
6. Explanation to grader	5	_____
7. Be able to hand-calculate the indirect, direct, and total effects from the a, b, & c' paths	5	_____
<b>Totals</b>	35	_____

---

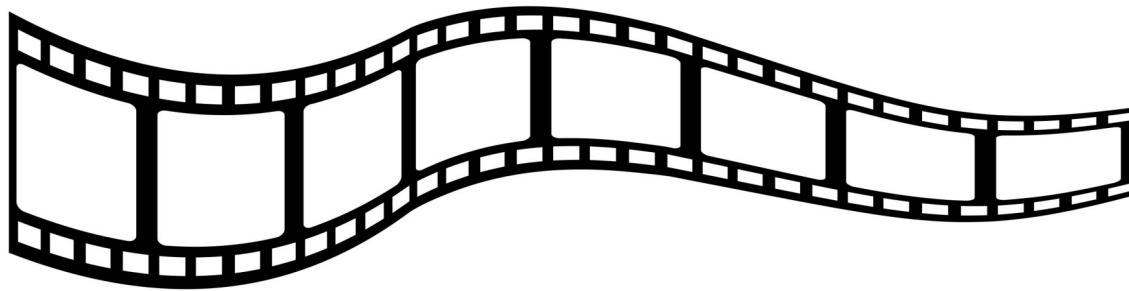


Figure 8.9: Image of a filmstrip

## 8.9 Bonus Track:

I find it useful to have script with the variables labeled merely by their role. Below, I quickly create an ModMeddemo\_df from the Lewis et al. [Lewis et al., 2017] simulated data and then run the analysis with the variables named in their roles.

```
library(tidyverse)
ModMedDemo_df <- rename(Lewis_df, X = GRMS, Y = MntlHlth, W = GRIcntlty,
M = DisEngmt)

library(lavaan)
set.seed(190505)
ModMedDemo <- '
#equations
M ~ a1*X + a2*W + a3*X:W
Y ~ c_p1*X + c_p2*W + c_p3*X:W + b*M

#intercepts
M ~ M.mean*1
Y ~ Y.mean*1

#means, variances of W for simple slopes
W ~ W.mean*1
W ~~ W.var*W

#index of moderated mediation, there will be an a and b path in the product
#if the a and/or b path is moderated, select the label that represents the moderation
imm := a3*b

#Note that we first create the indirect product, then add to it the product of the imm and
indirect.SDbelow := a1*b + imm*(W.mean - sqrt(W.var))
indirect.mean := a1*b + imm*(W.mean)
```

```

indirect.SDabove := a1*b + imm*(W.mean + sqrt(W.var))

#direct effect is also moderated so calculate with c_p1 + c_p3
direct.SDbelow := c_p1 + c_p3*(W.mean - sqrt(W.var))
direct.Smean := c_p1 + c_p3*(W.mean)
direct.SDabove := c_p1 + c_p3*(W.mean + sqrt(W.var))

#total effect
total.SDbelow := direct.SDbelow + indirect.SDbelow
total.mean := direct.Smean + indirect.mean
total.SDabove := direct.SDabove + indirect.SDabove
'

ModMedDemo_fit <- sem(ModMedDemo, data = ModMedDemo_df, se = "bootstrap", missing = 'fiml', bo

## Warning in lav_partable_vnames(FLAT, "ov.x", warn = TRUE): lavaan WARNING:
##   model syntax contains variance/covariance/intercept formulas
##   involving (an) exogenous variable(s): [W]; These variables will
##   now be treated as random introducing additional free parameters.
##   If you wish to treat those variables as fixed, remove these
##   formulas from the model syntax. Otherwise, consider adding the
##   fixed.x = FALSE option.

ModMed_FitSum <- summary(ModMedDemo_fit, standardized = TRUE, rsq=T, ci=TRUE)
ModMed_ParamEsts <- parameterEstimates(ModMedDemo_fit, boot.ci.type = "bca.simple", standardize = TRUE)

```

# References



# Bibliography

FACT SHEET: Anti-Asian Prejudice March 2020, a. URL <https://www.csusb.edu/sites/default/files/FACT%20SHEET-%20Anti-Asian%20Hate%202020%203.2.21.pdf>.

STOP AAPI HATE, b. URL <https://stopaaphate.org/>.

Hector Y. Adames, Nayeli Y. Chavez-Dueñas, and Maryam M. Jernigan. The fallacy of a raceless Latinidad: Action guidelines for centering Blackness in Latinx psychology. *Journal of Latinx Psychology*, 9(1):26–44, February 2021. ISSN 2578-8086. doi: 10.1037/lat0000179. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2020-69785-001&site=ehost-live>. Publisher: Educational Publishing Foundation.

Reuben M Baron and David A Kenny. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986. doi: 0022-3514/86. URL <https://www.sesp.org/files/The%20Moderator-Baron.pdf>.

Kenneth A. Bollen and Rick H. Hoyle. Perceived cohesion: A conceptual and empirical examination. *Social Forces*, 69(2):479–504, December 1990. ISSN 0037-7732. doi: 10.2307/2579670. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=1991-21226-001&site=ehost-live>. Publisher: University of North Carolina Press.

Barbara M. Byrne. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, Third Edition*. Taylor & Francis Group, London, UNITED KINGDOM, 2016. ISBN 978-1-317-63313-6. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4556523>.

Cristalis Capielo Rosario, Hector Y. Adames, Nayeli Y. Chavez-Dueñas, and Roberto Renteria. Acculturation Profiles of Central Florida Puerto Ricans: Examining the Influence of Skin Color, Perceived Ethnic-Racial Discrimination, and Neighborhood Ethnic-Racial Composition. *Journal of Cross-Cultural Psychology*, 50(4):556–576, May 2019. ISSN 0022-0221, 1552-5422. doi: 10.1177/0022022119835979. URL <http://journals.sagepub.com/doi/10.1177/0022022119835979>.

Charles S. Carver. You want to measure coping but your protocol's too long: Consider the Brief COPE. *International Journal of Behavioral Medicine*, 4(1):92–100, 1997. ISSN 1070-5503. doi: 10.1207/s15327558ijbm0401\_6. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=1999-13167-006&site=ehost-live>. Publisher: Lawrence Erlbaum.

- Jacob Cohen, P. Cohen, Stephen G. West, and Leona S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. LErlbaum Associates, Mahwah, N.J., 3rd ed. edition, 2003. ISBN 978-0-8058-2223-6.
- M. R. Cohen and E. Nagel. *An introduction to logic and scientific method*. Harcourt Brace, New York, 1934.
- Jose M. Cortina. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1):98–104, 1993. ISSN 0021-9010. doi: 10.1037/0021-9010.78.1.98. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.78.1.98>.
- Geoff Cumming. The New Statistics: Why and How. *Psychological Science*, 25(1):7–29, January 2014. ISSN 0956-7976. doi: 10.1177/0956797613504966. URL <https://doi.org/10.1177/0956797613504966>.
- Craig K. Enders. *Applied missing data analysis*. Guilford Press, New York, NY, 2010. ISBN 978-1-60623-639-0.
- Craig K. Enders. Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98:4–18, November 2017. ISSN 0005-7967. doi: 10.1016/j.brat.2016.11.008. Publisher: Elsevier Science.
- Andy P. Field. *Discovering statistics using R*. Sage, Thousand Oaks, California, 2012. ISBN 978-1-4462-0046-9.
- Andrew F. Hayes. *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach*. Guilford Publications, New York, UNITED STATES, 2018. ISBN 978-1-4625-3467-8. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=5109647>.
- Andrew F. Hayes. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guilford Publications, New York, UNITED STATES, 2022a. ISBN 978-1-4625-4905-4. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=6809031>.
- Andrew F. Hayes. More than one Mediator. In *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, pages 159–200. Guilford Publications, New York, UNITED STATES, 2022b. ISBN 978-1-4625-4905-4. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=6809031>.
- Sylvia Hurtado. Linking Diversity with the Educational and Civic Missions of Higher Education. *Review of Higher Education: Journal of the Association for the Study of Higher Education*, 30(2):185–196, 2007. ISSN 0162-5748. doi: 10.1353/rhe.2006.0070. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2006-23268-004&site=ehost-live>. Publisher: Johns Hopkins University Press.
- Sylvia Hurtado and Deborah Faye Carter. Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging. *Sociology of Education*, 70:324–345, October 1997. ISSN 00380407. doi: 10.2307/2673270. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=eue&AN=507591795&site=ehost-live>.

- Juliette M. Iacovino and Sherman A. James. Retaining Students of Color in Higher Education: Expanding Our Focus to Psychosocial Adjustment and Mental Health. In *The Crisis of Race in Higher Education: A Day of Discovery and Dialogue*, volume 19 of *Diversity in Higher Education*, pages 61–84. Emerald Group Publishing Limited, January 2016. ISBN 978-1-78635-710-6 978-1-78635-709-0. doi: 10.1108/S1479-364420160000019004. URL <https://doi.org/10.1108/S1479-364420160000019004>.
- Rajiv S. Jhangiani, I.-Chant A. Chiang, Carrie Cuttler, and Dana C. Leighton. *Research Methods in Psychology*. August 2019. ISBN 978-1-9991981-0-7. doi: 10.17605/OSF.IO/HF7DQ. URL <https://kpu.pressbooks.pub/psychmethods4e/>.
- K. G. Joreskog. Testing structural equation models. In Kenneth A. Bollen and J. Scott Long, editors, *Testing Structural Equation Models*. SAGE, February 1993. ISBN 978-0-8039-4507-4. Google-Books-ID: FvIxxeYDLx4C.
- Aycan Katitas. Getting Started with Multiple Imputation in R, 2019. URL <https://uvastatlab.github.io/2019/05/01/getting-started-with-multiple-imputation-in-r/>. Library Catalog: uvastatlab.github.io.
- Paul Kim. Yes, Asians and Asian Americans experience racism, March 2021a. URL <https://www.seattletimes.com/opinion/yes-asians-and-asian-americans-experience-racism/>. Section: Opinion.
- Paul Y. Kim. Guest Post: Anti-Asian Racism during the Pandemic: How Faculty on Christian Campuses Can Support Asian and Asian American Students, March 2021b. URL <https://christianscholars.com/guest-post-anti-asian-racism-during-the-pandemic-how-faculty-on-christian-campuses-can-support-asian-and-asian-american-students/>.
- Paul Youngbin Kim, Dana L. Kendall, and Hee-Sun Cheon. Racial microaggressions, cultural mistrust, and mental health outcomes among asian american college students. *American Journal of Orthopsychiatry*, 87(6):663–670, 2017. ISSN 0002-9432. doi: 10.1037/or0000203. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2016-39469-001&site=ehost-live>. Publisher: Educational Publishing Foundation.
- Rex B. Kline. The mediation myth. *Basic and Applied Social Psychology*, 37(4):202–213, July 2015. ISSN 0197-3533. doi: 10.1080/01973533.2015.1049349. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2015-36735-002&site=ehost-live>. Publisher: Taylor & Francis.
- Rex B. Kline. Data Preparation and Psychometrics Review (Chapter 4). In *Principles and practice of structural equation modeling*, pages 64–96. Guilford Publications, New York, UNITED STATES, 4th edition, 2016a. ISBN 978-1-4625-2336-8. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4000663>.
- Rex B. Kline. *Principles and practice of structural equation modeling*. Guilford Publications, New York, UNITED STATES, 4th edition, 2016b. ISBN 978-1-4625-2336-8. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4000663>.
- Jioni A. Lewis and Helen A. Neville. Construction and initial validation of the Gendered Racial Microaggressions Scale for Black women. *Journal of Counseling Psychology*, 62(2):289–302, April 2015. ISSN 0022-0167. doi: 10.1037/

- cou0000062. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2015-15500-004&site=ehost-live>. Publisher: American Psychological Association.
- Jioni A. Lewis, Marlene G. Williams, Erica J. Peppers, and Cecile A. Gadson. Applying intersectionality to explore the relations between gendered racism and health among Black women. *Journal of Counseling Psychology*, 64(5):475–486, October 2017. ISSN 0022-0167. doi: 10.1037/cou0000231. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2017-46640-003&site=ehost-live>. Publisher: American Psychological Association.
- Kaleea R. Lewis and Payal P. Shah. Black students' narratives of diversity and inclusion initiatives and the campus racial climate: An interest-convergence analysis. *Journal of Diversity in Higher Education*, October 2019. ISSN 1938-8926. doi: 10.1037/dhe0000147. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2019-59378-001&site=ehost-live>. Publisher: Educational Publishing Foundation.
- Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley, Hoboken, second edition. edition, 2002. ISBN 978-1-118-62586-6. URL <http://site.ebrary.com/id/10921256>.
- Todd D. Little, W. J. Howard, E. K. McConnell, and K. N. Stump. Missing data in large data projects: Two methods of missing data imputation when working with large data projects. *KUant Guides*, 011.3:10, 2008. URL <https://crmda.dept.ku.edu/guides/11.ImputationWithLargeDataSets/11.ImputationWithLargeDataSets.pdf>.
- R. Duncan Luce. Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, 46:1–26, 1995. ISSN 0066-4308. doi: 10.1146/annurev.ps.46.020195.000245. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=1995-19897-001&site=ehost-live>. Publisher: Annual Reviews.
- Brent Mallinckrodt, W. Todd Abraham, Meifen Wei, and Daniel W. Russell. Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology*, 53(3):372–378, July 2006. ISSN 0022-0167. doi: 10.1037/0022-0167.53.3.372. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2006-08780-009&site=ehost-live>. Publisher: American Psychological Association.
- Brent Mallinckrodt, Joseph R. Miles, and Jacob J. Levy. The scientist-practitioner-advocate model: Addressing contemporary training needs for social justice advocacy. *Training and Education in Professional Psychology*, 8(4):303–311, November 2014. ISSN 1931-3918. doi: 10.1037/tep0000045. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2014-25072-001&site=ehost-live>. Publisher: Educational Publishing Foundation.
- Della V. Mosley, Helen A. Neville, Nayeli Y. Chavez-Dueñas, Hector Y. Adames, Jioni A. Lewis, and Bryana H. French. Radical hope in revolting times: Proposing a culturally relevant psychological framework. *Social and Personality Psychology Compass*, 14(1), January 2020. ISSN 1751-9004. doi: 10.1111/spc3.12512. Publisher: Wiley-Blackwell Publishing Ltd.

- Della V. Mosley, Candice N. Hargons, Carolyn Meiller, Blanka Angyal, Paris Wheeler, Candice Davis, and Danelle Stevens-Watkins. Critical consciousness of anti-Black racism: A practical model to prevent and resist racial trauma. *Journal of Counseling Psychology*, 68(1):1–16, January 2021. ISSN 0022-0167. doi: 10.1037/cou0000430. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2020-20397-001&site=ehost-live>. Publisher: American Psychological Association.
- In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100, February 2003. ISSN 00222496. doi: 10.1016/S0022-2496(02)00028-7. URL <https://linkinghub.elsevier.com/retrieve/pii/S0022249602000287>.
- Kevin L. Nadal. The Racial and Ethnic Microaggressions Scale (REMS): Construction, reliability, and validity. *Journal of Counseling Psychology*, 58(4):470–480, October 2011. ISSN 0022-0167. doi: 10.1037/a0025193. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2011-19058-001&site=ehost-live>. Publisher: American Psychological Association.
- E. D. Niemark and W. K. Estes. *Stimulus sampling theory*. Holden-Day, San Francisco, CA, 1967.
- Mike C. Parent. Handling item-level missing data: Simpler is just as good. *The Counseling Psychologist*, 41(4):568–600, May 2013. ISSN 0011-0000, 1552-3861. doi: 10.1177/0011000012445176. URL <http://journals.sagepub.com/doi/10.1177/0011000012445176>.
- Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2000. ISBN 978-1-139-64936-0.
- Joseph Lee Rodgers. The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, 65(1):1–12, January 2010. ISSN 0003-066X. doi: 10.1037/a0018326. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2009-24989-001&site=ehost-live>.
- Yves Rosseel. The lavaan tutorial, January 2020. URL <http://lavaan.ugent.be/tutorial/tutorial.pdf>.
- Robert M Sellers, Stephanie A J Rowley, Tabbye M Chavous, J Nicole Shelton, and Mia A Smith. Multidimensional Inventory of Black Identity: A Preliminary Investigation of Reliability and Construct Validity. page 11.
- Klaas Sijtsma. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, 74(1):107–120, March 2009. ISSN 0033-3123. doi: 10.1007/s11336-008-9101-0. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792363/>.
- Anneliese Singh. Building a Counseling Psychology of Liberation: The Path Behind Us, Under Us, and Before Us. *The Counseling Psychologist*, 48(8):1109–1130, November 2020. ISSN 0011-0000, 1552-3861. doi: 10.1177/0011000020959007. URL <http://journals.sagepub.com/doi/10.1177/0011000020959007>.
- EF Stone-Romero and Patrick Rosopa. Research design options for testing mediation models and their implications for facets of validity. *Journal of Managerial Psychology*, 25:697–712, September 2010. doi: 10.1108/02683941011075256.

- Dawn M. Szymanski and Danielle Bissonette. Perceptions of the LGBTQ College Campus Climate Scale: Development and psychometric evaluation. *Journal of Homosexuality*, 67(10):1412–1428, August 2020. ISSN 0091-8369, 1540-3602. doi: 10.1080/00918369.2019.1591788. URL <https://www.tandfonline.com/doi/full/10.1080/00918369.2019.1591788>.
- Paolo Toffanin. Multiple-mediator analysis with lavaan, May 2017. URL <https://paolotoffanin.wordpress.com/2017/05/06/multiple-mediator-analysis-with-lavaan/>.
- Clairice T. Veit and John E. Ware. The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology*, 51(5):730–742, October 1983. ISSN 0022-006X. doi: 10.1037/0022-006X.51.5.730. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=1984-02935-001&site=ehost-live>. Publisher: American Psychological Association.
- John E. Ware, Mark Kosinski, Martha S. Bayliss, and Colleen A. McHorney. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the medical outcomes study. *Medical Care*, 33(4, Suppl):264–279, April 1995. ISSN 0025-7079. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=1995-35535-001&site=ehost-live>. Publisher: Lippincott Williams & Wilkins.