HW 2 and Lab 2

Nicky Wakim

2025-02-12

Homework 2

A small word on the homework

- Mostly good work!
- Main note: Please look at the solutions to make sure you have the correct beta's and interpretations when we work with categorical variables!!

Lab 2

A note from me

- I know the lab instructions are wordy
- This class is really about the technical ("objective") skills of regression
  - But in order to responsibly practice statistics, you need to critically think about the **subjective** choices you make
- And I'm really trying to lay out my thought process in the labs so that you have some idea of the subjective choices that I'm kinda restricting us to
  - And that's really just bc you're learning A LOT in this class
  - So taking on extra learning objectives would be overwhelming

Other overall issues

- No need to load the codebook into R!!!
  - Codebooks are typically opened in excel and will give you extra information on the variables
- You gotta show all your code!
  - If you got points off for not showing any code, resubmit with the code showing and I'll give you credit back
- Be careful when making assumptions about the data
  - Example: someone created a cisgender variable by seeing if SAB was the same as gender identity
    - I would be wary of that - definitions of cis and trans are highly personal - only use and refer to participants as they self-identify
- Do **not** immediately make age categories!! It is important to look at age (numeric) vs. IAT
  - Why pixelate your data?? We only do it if we need to (aka age as numeric is NOT linear with IAT score)

3.1: What is our target population?

- This is an important thing to flag as you analyze your results and interpret them for an audience

- We restricted our population to the US
- Harvard says the test is only for individuals 18+ years old
- Test takers need access to the internet and a computer (or phone?)

- Another thought
  - Sometimes your target population defines your sample
  - Other times your sample defines your target population
- Here we have a convenience sample, with specific restrictions and accessibility
  - That means the population that we can generalize to is limited to those restrictions and accessibility!!
- **We need to discuss these limitations when we present these results to the world!**

3.2 Restrict your analysis to 1 outcome and 9 possible covariates/predictors

Needed to pick the variable from your research question + 2 others (or 3 if you chose a different variable in your research question)

1. Explicit anti-fat bias (`att7`)
2. Self-perception of weight (`iam_001`)
3. Fat group identity (`identfat_001`)
4. Thin group identity (`identthen_001`)
5. Controllability of weight of others (`controlother_001`)
6. Controllability of weight of yourself (`controlyou_001`)
7. Awareness of societal standards (`mostpref_001`)
8. Internalization of societal standards (`important_001`)

Needed to include all 4 demographic variables:

1. Age (we need to construct from `birthmonth`, `birthyear`, `testmonth`, and `testyear`)
2. Race (`raceomb_002` **or** `raceombmulti`)
3. Ethnicity (`ethnicityomb`)
4. Sex assigned at birth (`birthSex`)

Please pick only 2 additional variables:

1. Education (`edu_14`)
2. Gender (`genderIdentity`)
3. Self-reported BMI (through self-reported height and weight)
4. Political identity
5. Religion

3.2 Restrict your analysis to 1 outcome and 9 possible covariates/predictors

- Start by loading the data

```
load(file = here("../Project/data/iat_data.rda"))
iat_2021 = iat_2021 %>%
  select(IAT_score = D_biep.Thin_Good_all,
         att7, iam_001, identfat_001,
         myweight_002, myheight_002,
         identthin_001, controlother_001,
         controlyou_001, mostpref_001,
         important_001,
         birthmonth, birthyear, month, year,
         raceomb_002, raceombmulti, ethnicityomb,
         edu, edu_14,
         genderIdentity,
         birthSex) %>%
  drop_na()
```

3.3: Manipulating variables that are coded as numeric variables

- No need to make plots here (that was just part of my example)
  - Plots and tables are a good way to check that you accomplished the correct translation
- Giving the levels order:

```r
iat_2021 = iat_2021 %>% mutate(iam_001_f = case_match(iam_001,
                                       7 ~ "Very overweight",
                                       6 ~ "Moderately overweight",
                                       5 ~ "Slightly overweight",
                                       4 ~ "Neither underweight nor underweight",
                                       3 ~ "Slightly underweight",
                                       2 ~ "Moderately underweight",
                                       1 ~ "Very underweight",
                                       .default = NA) %>%
                 factor(levels = c("Very underweight", # Assigns the level order!
                                   "Moderately underweight",
                                   "Slightly underweight",
                                   "Neither underweight nor underweight",
                                   "Slightly overweight",
                                   "Moderately overweight",
                                   "Very overweight")))
```

3.3: Manipulating variables that are coded as numeric variables

- Now when we print a table, we can see them in order

```
iat_2021 %>%
  dplyr::select(iam_001_f) %>%
  tbl_summary()
```

Characteristic	N = 242,762[1]
iam_001_f	
Very underweight	1,341 (0.6%)
Moderately underweight	5,436 (2.2%)
Slightly underweight	17,224 (7.1%)
Neither underweight nor underweight	106,836 (44%)
Slightly overweight	65,418 (27%)
Moderately overweight	32,259 (13%)
Very overweight	14,248 (5.9%)

[1] n (%)

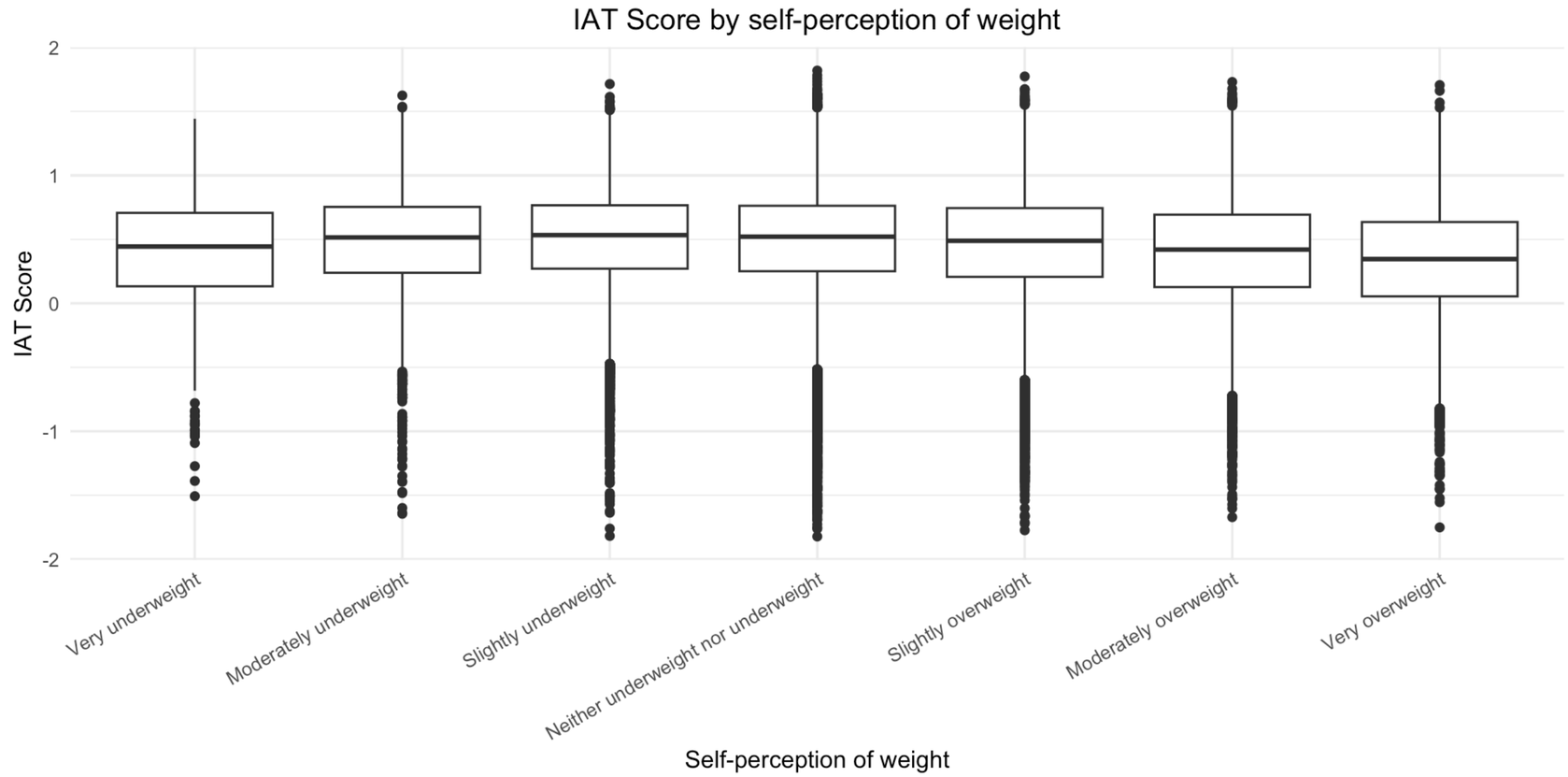
3.5 If you chose BMI, create the variable

- If you worked with BMI, please make sure you followed the help page!
- Please come double check with me that you are creating it correctly!

4.3 Bivariate exploratory data analysis

- You only needed to create one plot!!
- My research question: Is self-perception of weight associated with IAT score?

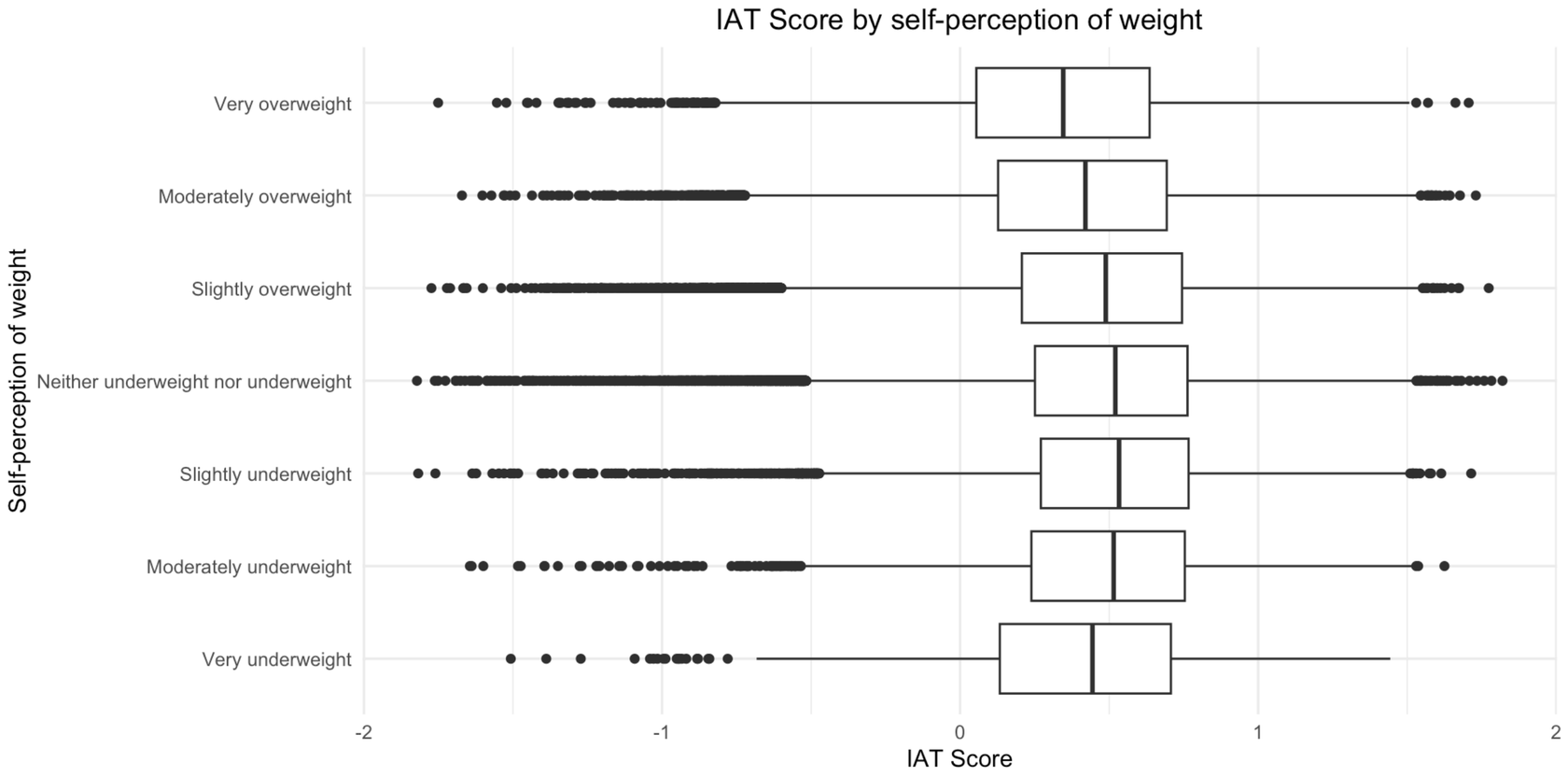
▶ How I made the plot



4.3 Bivariate exploratory data analysis

- You only needed to create one plot!!
- My research question: Is self-perception of weight associated with IAT score?

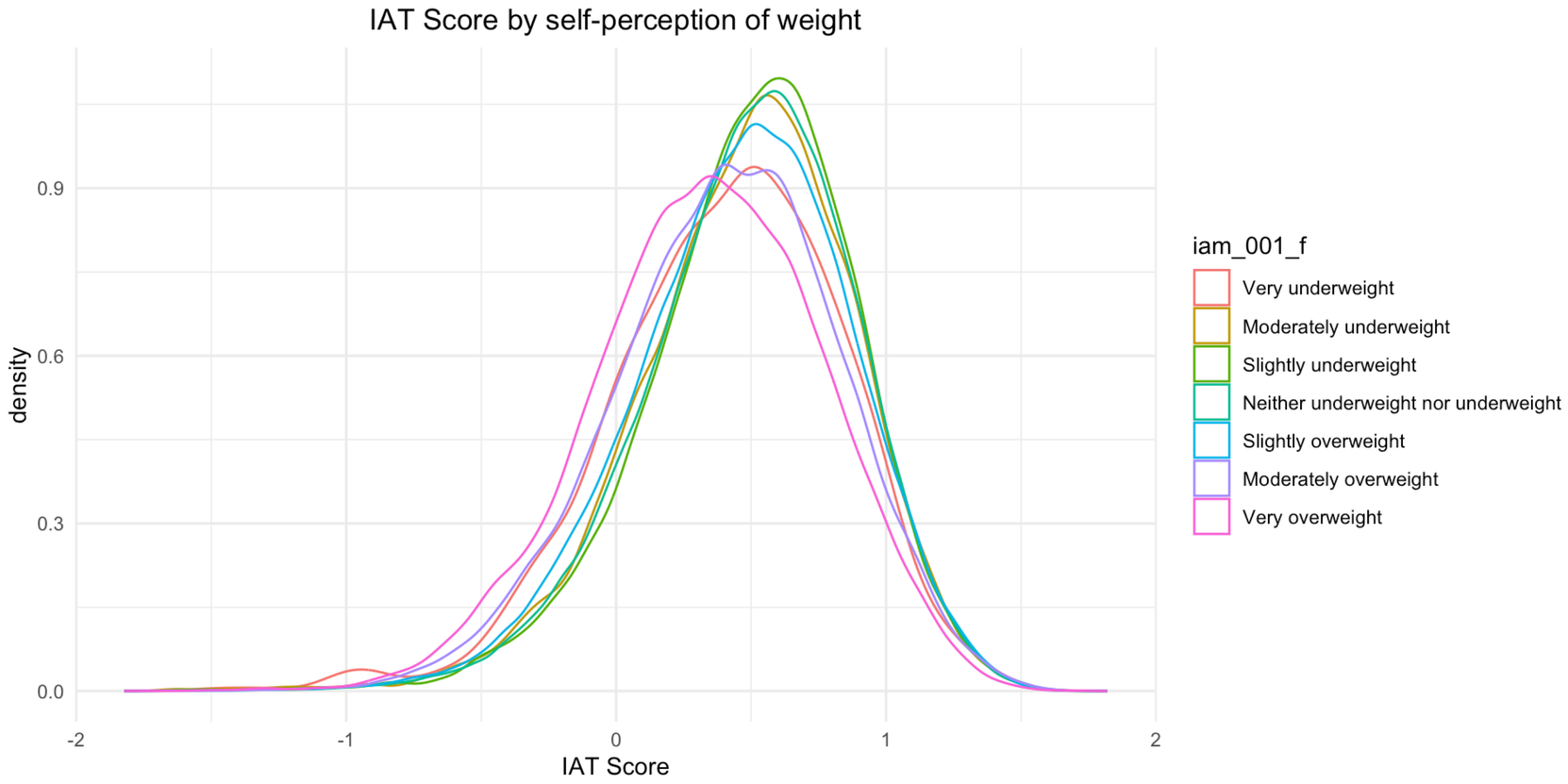
▶ How I made the plot



4.3 Bivariate exploratory data analysis

- You only needed to create one plot!!
- My research question: Is self-perception of weight associated with IAT score?

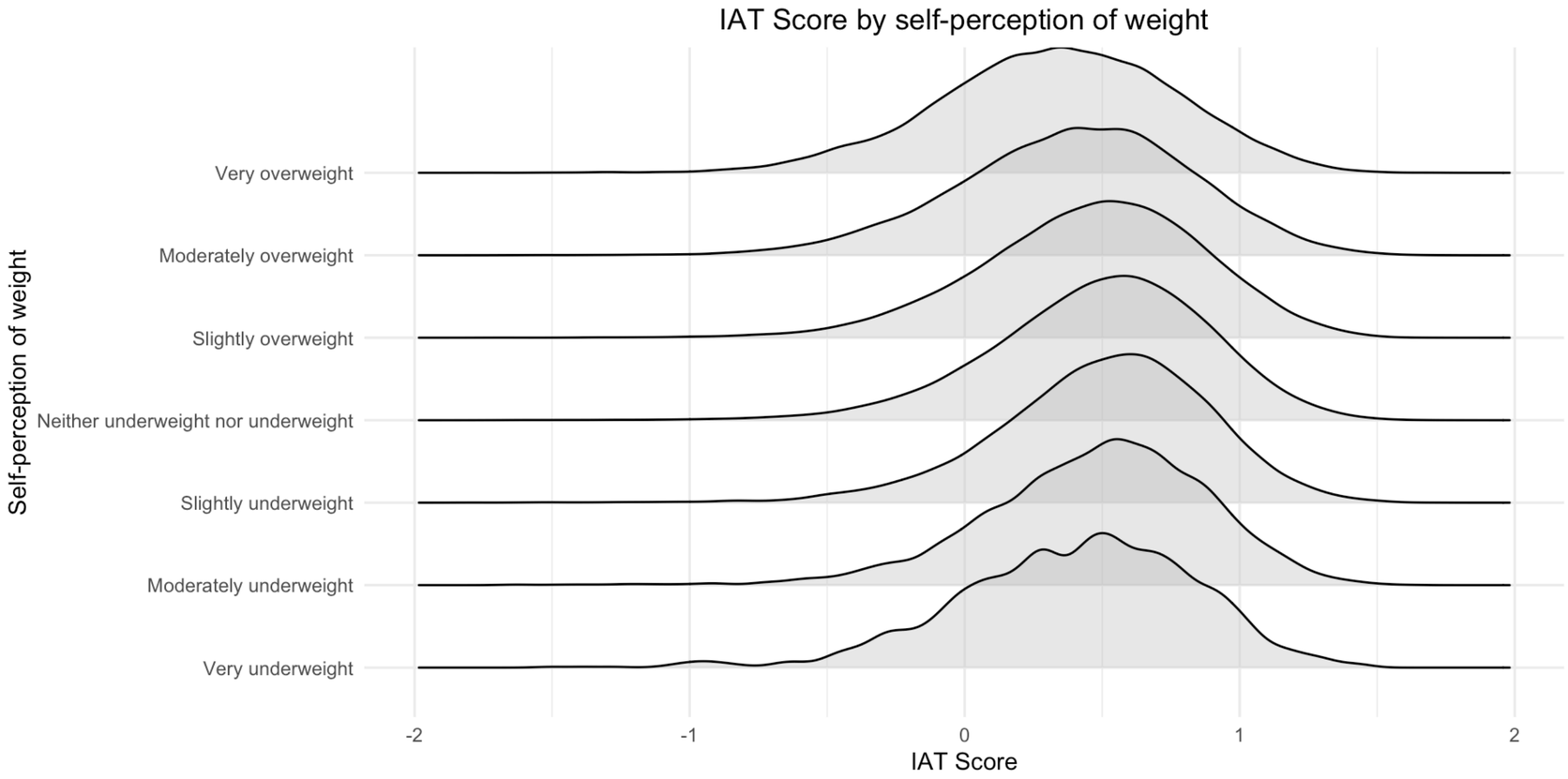
▶ How I made the plot



4.3 Bivariate exploratory data analysis

- You only needed to create one plot!!
- My research question: Is self-perception of weight associated with IAT score?

▶ How I made the plot



Multi-selection/multi-response variables

Multi-response/multi-selection variables: `raceombmulti` and `genderIdentity`

- There is extra data management skills that we need to address these
- Let's walk through categorical variables that have multiple selections
- Another note: I find that the race variable is still lacking (where is my MENA representation??)
  - MENA = Middle Eastern and North African

4 Approaches to Multiple-Race Questions from We All Count

- This method works for any multi-level variable

Why, when, and for whom we would use a multiple-race question are all really, really important questions that I'll talk a little bit about at the end of the article, but before that, I want to talk about a few different ways we can analyze these questions.

Let's say our multiple-race question looks like this:

What category describes you? (Select all that apply):

American Indian
Black
LatinX
White

What makes this a multiple-race question is the "select all that apply" next to the title. Some respondents will likely check two or more of those boxes.

Let me be clear; this example is not a suggestion. How the question is phrased, the use of "American Indian", the use of "Black", the use of "Latinx", the use of "White", the lack of other options, etc all need to be considered, but for right now, I want you to focus on the fact that there are four options and you can choose more than one.

If I were handed the results from this survey and asked to analyze them, I would think about four little symbols which represent four different ways we might think about this data:

_

&

x

/

– : The Hyphen Approach

The first symbol is the hyphen: – . If I'm analyzing the data with a "hyphen mentality", when people select multiple races, they become an entirely new category:

Black-LatinX or American Indian-White, for example.

These categories are not 50% one thing and 50% the other; they are a unique category reflecting an approach that considers being Black-White its own kind of experience. I will end up with as many categories as there are combinations selected. I might use this approach if I'm asking a race question to explore the unique experience of what I consider distinct categories.

There are some interesting things to think about with the hyphen approach, like primary vs. secondary racial identities (i.e., is White-LatinX the same thing as LatinX-White?) and limits on the number of possible combinations. This approach (and some others we'll get into below) can lead to very small sub-group sizes and low levels of certainty for those groups due to the nature of quantitative data, so it would be important to think about that and the trade-off between inclusive survey questions and creating meaning for the smallest groups your categories lead to.

& : The Ampersand Approach

I could instead think about respondents with multiple races like this:

Black & LatinX

Rather than a unique category, I think about the racial backgrounds that make up this person's identity as separate factors with their own impact on whatever I'm researching. Let's say I am looking at how race affects someone's likelihood of being misdiagnosed in the emergency room. If I think being Black has a particular effect on the likelihood of misdiagnosis and being LatinX has another, I might want to keep them separate as components of someone's experience.

Analyzing like this means you are more interested in the component than the person's holistic experience, and you can quickly get into situations where individuals are counted in multiple columns in the table below:

	American Indian	Black	LatinX	White
Respondent 1				X
Respondent 2		X		
Respondent 3	X		X	
Respondent 4	X		X	X
Respondent 5		X		X
Etc…				

When I talk about results for "White" racial component, I would be talking about respondents 1, 4, and 5. When I talk about "Black" in my report, I would be talking about respondent 2, but also respondent 5 again.

The ampersand approach comes from a different place than the hyphen approach. Its worldview about what race is, what having more than one means, and how race should be treated in the analysis is not the same. It comes with different technical challenges, like counting/weighting issues. It can even lead to novel approaches like thinking about each person's race as a pie chart of racial components and standardizing the pieces of the pie by however many categories you end up with. Instead of these results:

Respondent 1: White

Respondent 2: Black

Respondent 4: American Indian & LatinX & White

We'd get something like this:

Respondent 1: White & White & White

Respondent 2: Black & Black & Black

Respondent 3: American Indian & LatinX & White

Okay, we're getting crazy here, but I'm trying to show that there are no limits to how we can think about "race" and, therefore, no limits to how we might need to analyze it. Anyway…

X : The Multiplicative (Intersectional) Approach.

The ‘x’ approach is a very unconventional and off-label use of an idea called intersectionality. In a nutshell, intersectionality points out that sometimes the effect of different elements of someone’s identity or experience shouldn’t be siloed (as in the & approach) nor combined (as in the – approach) but rather multiplied. If, for example, there are barriers to employment facing black people and there are barriers to employment facing women, the experience of someone who is both of those things is better described like this:

black x woman = % barrier to employment

Black women may be experiencing a compounded barrier due to the multiplicative interactions of the unique oppression they face due to more than one facet of who they are. Put more simply; Black women might face more barriers more severe than Black people or women! Well, this might blow your mind, but we can do the same thing within one facet like race:

Black x American Indian = % barrier to employment

If we use the x approach, we are interested not in the effects of multiple race components alone (like in the & approach), but rather in the effect they have together (and even on each other).

The x approach has the benefit of mathematically reflecting a perspective that says the reality of being multiple races is more than the sum of two separate things. Whenever we adopt an approach to these questions, we're saying, "for the purposes of this model, we think "race" works like this in the real world". And sometimes it doesn't work like that. If we return to a question like "likelihood of misdiagnosis", and we think that being Black increases your likelihood of misdiagnosis and being White reduces it, we could do the math with the x approach:

white = low likelihood

black = high likelihood

Black x White = medium likelihood? Maybe, maybe not… it depends on what we think "race" is a proxy for…

I suppose in some contexts being 50% White brings 50% of the privilege, but not usually and certainly not historically. Especially not if you are measuring racism or oppression. Exclusionary and racist systems are generally very binary (you either "are" or you "aren't"), expressly so they can exclude the greatest number of people. Speaking of binary…

/ : The Slash (binary) Approach

One of the important things to think about when crafting multiple-race questions is the use of options like “mixed-race” or “other”. It’s a very fraught decision. There are many important things to discuss regarding how those category labels and ones like them feel to survey respondents, but here I want to discuss the analysis ramifications.

Sometimes the question we're trying to answer involving multiple races is about the difference between having one, zero, or multiple racial identities. Sometimes the question is about whether we are at all part of a certain racial group or not. We could analyze our responses into binary categories, and that's where the slash comes in:

"White" only / Everybody else

or

"LatinX" / Everybody Else
(This one might include people who selected "LatinX" as well as other options)

or

Single Race / Multiple Races

or

One or more of our survey options / "Other"

Each of these breakdowns reflects a different thing we want to know. Maybe we're trying to measure against the experience of people who identify as "White" exclusively. Perhaps we're trying to answer questions about the experience of having an identity made up of multiple races. Maybe we're interested in how well our collection tool captured how people categorize themselves.

One advantage to this approach is its highly aggregated categories which can be helpful if you have small group sizes and want some more confident, if less specific, answers. The major downside is that this approach is only appropriate to a certain number of questions, and these might not be the ones you have.

Ooooof... what should I do?!

That there are many ways to analyze multiple-race questions points to the fact that there are many ways to use the idea of "race" in your work. Some of these ways are good. Some are bad. When you adopt and craft multiple-race questions, how you analyze them, and how you report them reveals your own, your project's, or your organization's worldview of "race". Even if you don't make this choice consciously or deliberately, a choice is being made.

If you are like me, thinking about concrete, practical decisions (like in which of these four ways do we want to approach this data) can lead to more constructive and efficient conceptual meetings about a crucial yet downright Gordian subject like "race". Let's be clear, questions like "do we really want to ask about race?", "Can you be more than one race? If so, how many?", "What is the difference between race/culture/ethnicity/nationality/background?", "Is this a bad proxy for something else?" or "are we perpetuating racism by even acknowledging "race" as a thing?!" are very, very important.

Next, questions like: "what question are we trying to answer?", "what are our priorities around inclusivity versus certainty?", "is what we're going to do with this data clear from how we're getting it?", or "how do our respondents feel about this question?", etc., are necessary to design your project effectively.

I know it's not easy and that many people using multiple-race questions are doing so because they have been asked or required to do so. Thinking about – vs. & vs. x vs. / has helped me a lot when I have to move on from the theory and put these questions into the field or generate results for a report. It's also invaluable when explaining my work to others. Matching the how to the why tends to improve the how and clarify the why.

Final notes

- For now, I suggest the binary approach!
  - This is the perfect level of pushing ourselves coding wise and thinking critically about these multi-response variables
- Take a look at this article: https://doi.org/10.1016/j.socscimed.2017.12.026
  - It gets into some of the considerations and uses of intersectionality in analyses