STATISTICAL HORIZONS

Psychometrics

Matthew Diemer, Ph.D.

Upcoming Seminar:
November 18-20, 2021, Remote Seminar

MODERATION, FAIRNESS & BIAS: MIMIC MODELS

CONSTRUCT BIAS: APPLES, ORANGES & PIGGLY WIGGLY

Imagine a scale at PriceChopper, Star Market, Piggly Wiggly, or your favorite grocery store.

When you (omnisciently) put 10 lbs of apples on that scale, it reads 10 lbs.

When you (omnisciently) put 10 lbs of oranges on that scale, it reads 7 lbs (?)

This scale exhibits *construct bias* in how it measures oranges. From the perspective of the scholar Janet Helms, oranges are not measured with fairness (putting aside what experiences & perspectives of being an orange lead to this).

So: If the weight of oranges is used to predict some other thing -> amount of juice extracted from the fruit, for ex. -> then we will not have fairness in how weight measurements are used

This could also be characterized as ***predictive bias***

Or: ***"biased tests yield score variance that reflects construct-irrelevant group differences, while good tests yield score variance that reflects construct- relevant group differences (e.g., high understanding of math concepts or high levels of experienced anxiety)." [quoting previous student's HW commentary]***

i.e., "random error variance or factors that are wholly unrelated to individual differences in the trait measured by the test" (Frisby, 1999, p. 264). -> why would gender predict math self-efficacy beliefs?

HID (HELMS INDIVIDUAL-DIFFERENCES) FAIRNESS MODEL

Summary: Psychological factors associated with race/ethnicity cause differences in test scores.

Race/ethnicity cannot be manipulated and is not a "variable." (Janet Helms – she/her - argues race is only a social construct, not a biological reality, yet also has profound social implications.)

What are these psychological factors? Examples include: Stereotype threat, impact(s) that feeling like a test is biased against you/your group, Immersion (pro-Black and anti-White attitudes) racial identity status

Furr & Bacharach, p. 304: Mechanical aptitude scores determined by mech. aptitude for males, yet by mech aptitude AND stereotype threat for females

Note: F & B and others characterize this as construct bias & Helms characterizes this as 'fairness' -> similar, but not the same

HID (HELMS INDIVIDUAL-DIFFERENCES) FAIRNESS MODEL (II)

Using race/ethnicity group membership (e.g., White vs African American) is only a **rough proxy** for these racial/ethnic psychological factors

If we include race-based psychological factors in our models, then we do more to achieve fairness in testing (& we pay more attention to consequential validity)

Ex: Enter Immersion status in 1st step of regression, then racial group in 2nd

OR: -.23 correlation btw Racial Identity (RI) & CAKS (p. 853, Table 1) -> 'Construct irrelevant-variance' in CAKS scores.'

Because we have evidence of construct-irrelevant variance in a measure (CAKS, mechanical aptitude – for females was aptitude + stereotype threat), then we have construct bias -> or, (un)fairness

A 'VALIDITY STRATEGY' TO ADDRESS FAIRNESS

- **MIMIC models:** I suggest as one of the "pragmatic strategies for identifying or removing unfairness from individual test takers' scores if construct-irrelevant variance is discovered" (Helms, 2006, p. 846)
  -> if that construct-irrelevant variance is *group-based measurement error*
  -> this is a way of assessing *construct bias*, using the terms of the F & B readings
- i.e., "random error variance or factors that are wholly unrelated to individual differences in the trait measured by the test" (Frisby, 1999, p. 264).
- MIMIC models would identify items that exhibit bias, in that they function differently across groups, one dimension of test fairness
- On the other hand, MIMIC models reify racial/ethnic categorization, and would only address issues of fairness inasmuch as categorizing race captures, or measures, dimensions of racial/ethnic socialization, cultural practice, etc., that contribute to group differences on tests (from Helms' perspective)
  - Further, race as category fails to capture structural racism, inequitable policing, anti-Blackness
  - MIMIC models would only measure racial/ethnic differences by proxy, via categorizing race/ethnicity
  - MIMICs do not capture the "racial or cultural psychological attribute" (e.g., stereotype threat, Immersion) that contribute construct-irrelevant variance to scores
- This is therefore more of a 'validity strategy' than a 'fairness strategy' to use the terminology from the Helms (2006) Table 4 (see examples 1-4 under 'validity strategy' column heading)

MIMIC MODELS: CONCEPTUAL BASIS

In education & in psychology, we often want to answer questions like the following:

- How do background characteristics (also called exogenous covariates, which are traditionally 0 or 1 in the MIMIC case) relate to differences on latent factors or observed indicators?
- Also stated: How does an exogenous covariate predict individual differences on a latent construct? How does an exogenous covariate predict responses to an individual item?
  - For ex: How does gender predict scores on an expressive language measure? How does gender predict responses to the individual items that comprise the expressive language measure (i.e., are the items biased?)
  - For ex: How do "low achievers" and "high achievers" differ when measuring X and Y?

MIMIC [Multiple Indicators and Multiple Causes] models – CFAs with exogenous covariate(s) – are well-suited to answer these questions

- To detect latent mean differences (which, because measurement error is parceled out, are more precise comparisons than t-tests of observed means) as well as differential item functioning (does group membership predict differential responding to an item)
  - Preview: Differential Item Functioning, or DIF, is an important aspect of Item Response Theory (IRT)

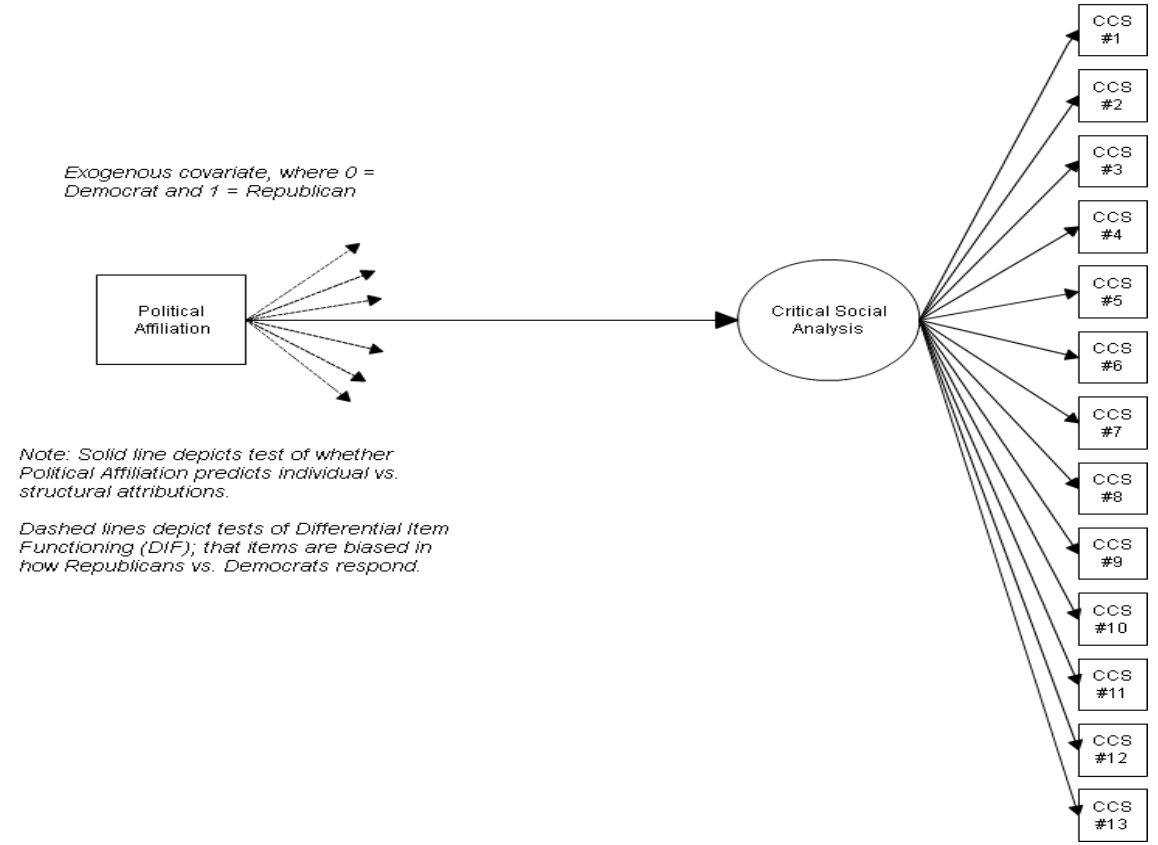
Student HW commentary quote: "A MIMIC model is used to address group differences in a measurement model and to test whether each latent factor is measured in the same way across groups. Then, if it is not, the model is re-specified to account for the group difference, which improves the structural model."

SAMPLE MIMIC

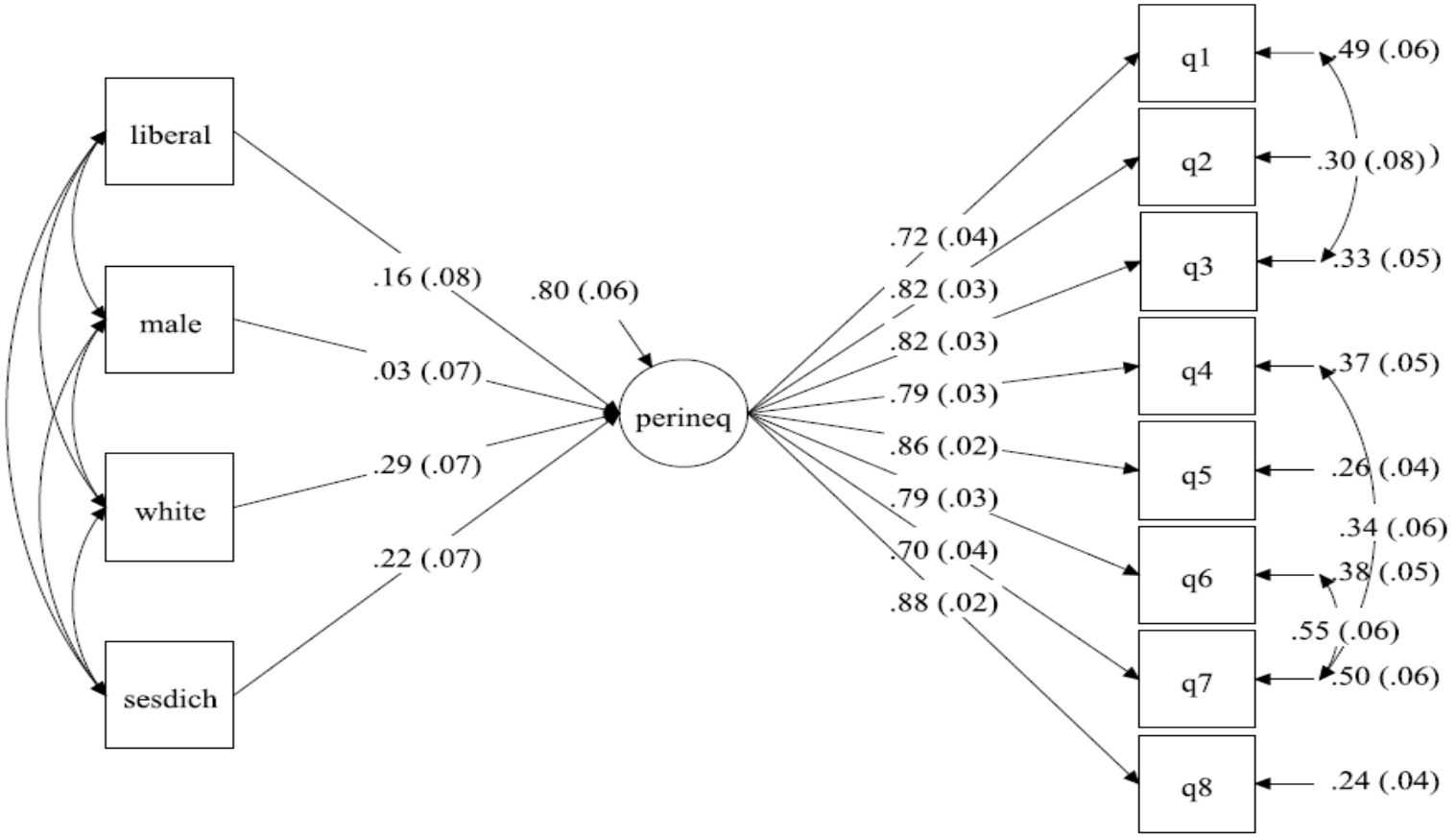
Note: We are used to seeing circles predict boxes (on the R side of this diagram); here we *instead* have boxes predicting circles on L side of diagram [***not*** *LHS/RHS equation*]

Note also: Paths from exogenous covariates to indicators only suggested here…

Figure 1: Testing Whether Political Affiliation Predicts Critical Social Analysis via a MIMIC Model



MIMIC: EMPIRICAL ILLUSTRATION



Fit Indices:
Chi Square- 60.415
RMSEA- .05
CFI- .985
TLI- .981
SRMR- .031
Path Male is NS

TECHNICAL SPECIFICATIONS OF MIMICS

- MIMIC models are a way of testing moderation – a moderator variable "alters the direction or strength of the relation between a predictor and an outcome" (Frazier, Tix & Barron, 2004).
  - Here, the moderating variable is group membership of some kind -> does group membership predict latent mean differences of differential item functioning (DIF)?
- SEM = very powerful & flexible analytic framework; MIMIC models = an SEM analysis that capitalizes on flexibility
- MIMIC models (traditionally) use dichotomous exogenous covariates → dummy code or dichotomize non-dichotomous variables (example later in these slides)
- MIMIC models are carried out with the entire sample, in contrast to multi-group CFA , which fits the same CFA model to each group, separately, & is reviewed later in this course
  - The 'USEOBSERVATIONS' & 'GROUPING' commands, which are specific to multi-sample CFAs -> are NOT used with MIMIC models
- Although MIMIC models are a variant of CFA models, there are no special considerations with regard to identification & estimation for MIMIC models
  - Same ULI or UVI identification strategies & same estimators (e.g., MLR, WLSMV, etc.) are used

WHAT DO MIMICS ESTABLISH?

Strong Invariance (a.k.a. Scalar invariance) -> presuppose configural (same 'configuration' of boxes to circles') & metric (same 'metric' of loadings across groups') invariance

Unstandardized intercepts between groups are constrained to be equal.

Two people from different groups with the same level on a certain factor have the same score on a given indicator

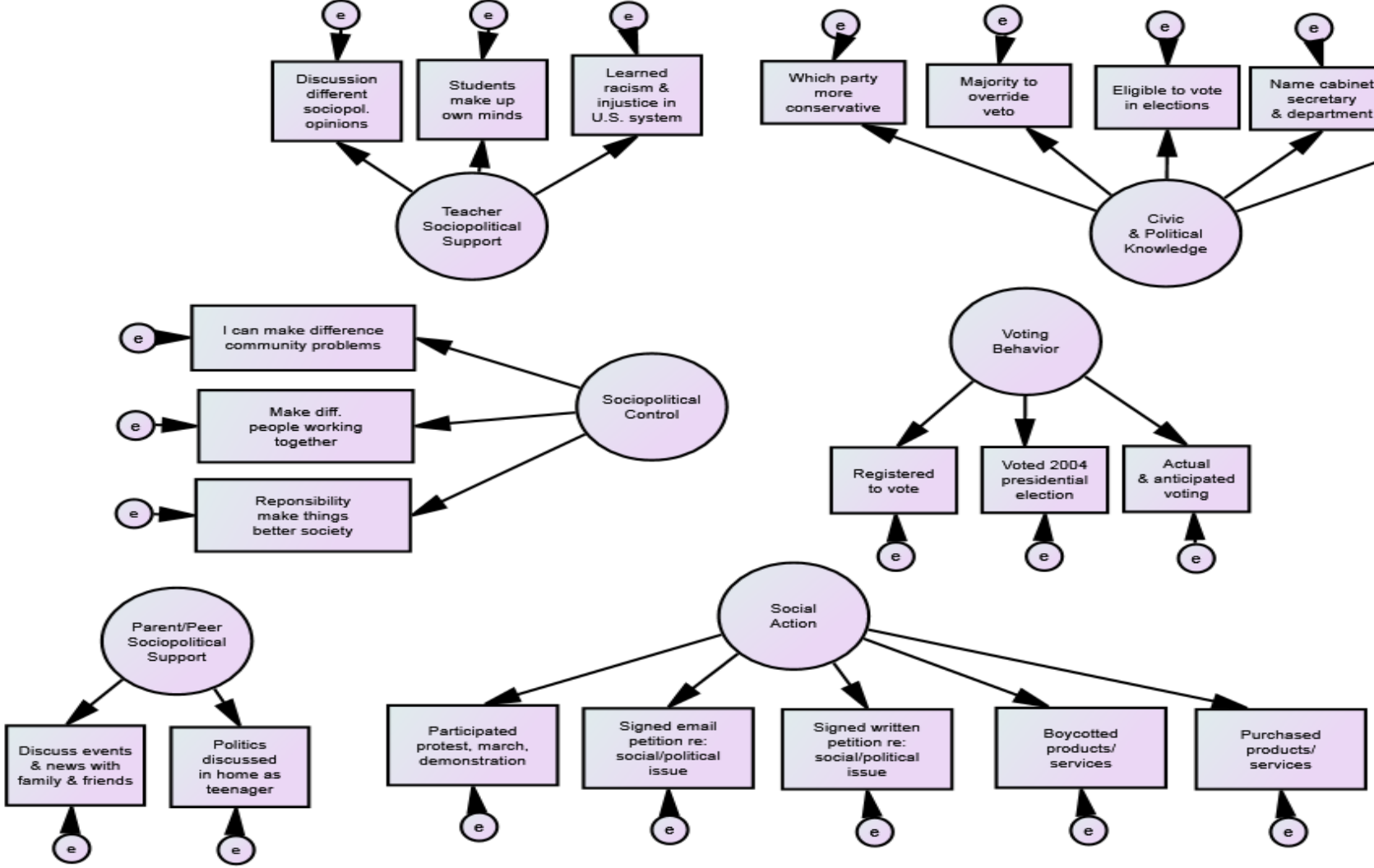
Scalar Invariance: Practical example: men and women with equal levels of depression self-report the same amount of binge eating

Vs Scalar variance:

At the same level of depression, women binge-eat twice per week and men binge-eat 4 times per week, even though the slopes (increases as depression increases) may be the same.

STEPS TO FIT A MIMIC MODEL

1. Finalize well-fitting CFA
2. Model exogenous covariate (s) – note the covariates do not correlate with each other & remain totally exogenous – with paths going from covariate to each latent variable
   a. If not pre-existing, covariates can be created in SPSS or in M+ [both reviewed later]
   b. What do exog covariate -> latent paths test?
3. Take note of significant latent mean differences detected & whether fit improves/decreases
4. Then, inspect modification indices (MIs) for items predicted by the exogenous covariate
   a. E.g. [having voted or anticipating voting] Q41_42CO ON AGE_R [dichotomous age covariate, where 0 = 15-20 and 1 = 21-24; 'split' here is theoretically informed in that older group was age-eligible to vote in preceding presidential election]
   b. *** Here is the one time that you let the modification indices lead you toward model respecification** (per Kline, Raykov, Muthén, Kaplan, others)*
5. *OR* use your own judgment of "suspects" – items that may be predicted by your covariate
6. Add suggested paths from covariate to items & inspect significance & model fit changes
   a. Significant item ON covariate relationships are evidence of DIF
   b. These DIFs can be retained in the structural model ('step two'], to *adjust structural model for group differences in the measurement of biased items*



CFA model of Diemer & Li, 2011

Chi-squre (264.637, P < .05)
RMSEA (.03)
CFI (.95) & TLI (.96)
WRMR [for what it's worth!] (.95)

MAIN TASK & SANDBOX, OPTION #1

- **CCS dataset, CFA model (use previous files)**
  - See Folder for SPSS and .dat files; inspect SPSS file to identify candidates for exog. covariate(s)
  - *Note this task corresponds to HW#3*
- **You know the CFA structure of the CCS, for some constructs, by now.**
  - Does fit improve when you proceed from CFA to the MIMIC model?

- **Following steps reviewed above, interpret fit of (a) CFA vs. (b) MIMIC model**
  - Do the covariates predict any latent mean differences? MODINDICES suggestions? DIF?
  - Interpret what any significant relationships mean
- **Try dichotomizing other variables from the CCS dataset you are working with. Add them as 'replacement' or additional covariates (recall you can have multiple exogenous covariates, but they generally are not freely correlated with each other)**
  - See later slides ahead for how to do so, with the PSID as an example

SUMMARY

Helms' Individual Differences (HID) model emphasizes that racial/ethnic category is a very coarse way of measuring race/ethnicity

I.e., racial/ethnic group is only a *proxy* for racial/ethnic identity status or regard

Fairness -> Thinking about the social consequences of psychometrics (i.e., consequential validity) AND statistical techniques to identify and model how race/ethnicity contributes to measurement

Bias -> MIMIC models simple, yet effective, strategy to detect and control for items that may be biased

Also afford testing for item bias while statistically adjusting for latent mean differences

Note: This does not establish Measurement Invariance (MI), a more intensive & stepwise process

One point of synthesis btw HID & MIMICs: Measure racial identity status and split participants into two groups, using mean or median splits, into "low" and "high" levels of Immersion racial identity status (or, stereotype threat – perhaps measured physiologically)

OR: high & low scores on the MIBI -> perhaps split on private regard or Nationalist profile

RECAP: GUIDING POINTS TO PROBE BIAS

- MIMICS sound intimidating…but as we have learned, are not that complicated to estimate and interpret
- An efficient strategy to detect bias, either in latent means (levels) or DIF (items biased across groups)
  - More precise than ANOVAs bc measurement error parceled out
  - Also: Easier & require lower N than full measurement invariance (MI) testing ->
  - "unlike multigroup factor analysis…, several covariates can be incorporated into the MIMIC model without subdividing the sample" (Gallo et al., 1994, p. 252).
- Steps:
  - [1] CFA, [2] latents ON exogenous covariate, [3] MODINDICES, [4] items ON covariate(s)

- Once measurement bias is detected, it can then be controlled in structural models ('step two')
  - This capacity to adjust SR models for measurement bias is important, yet under-utilized

- Limitations of MIMICs:
  - Group membership is modeled as a dichotomy, when its more than that -> can oversimplify racial/ethnic identification, intersectional identities, etc into a 0 or 1
  - Cannot declare measurement invariance, formally, with MIMICs