Latent variables:
What are they? Why is they?

Briana Mezuk, PhD
MCUAAAR AnC
13 December 2023

What does it mean for a variable to be "latent"?

- "Latent" refers to the idea that the variable is **not directly observed** (that is, **it is not measured** in your dataset) - and **may not even be observable** - but is **theorized** to exist.

"**I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description [of obscenity]**; and perhaps I could never succeed in intelligibly doing so. **But I know it when I see it**, and the motion picture involved in this case is not that."

-SCOTUS in Jacobellis vs. Ohio (1963)

What does it mean for a variable to be "latent"?

- The **characteristics** of latent variables (T) are **"inferred"** (i.e., quantified or extracted, λ_1, λ_2) based on the **correlation between >1 imperfectly-measured** (e_1, e_2…) but **observed "indicators"** (x_1, x_2…) of said latent variable.
- The observed indicators are **assumed** to be correlated with each other **because** they are "proxies" of the latent variable.
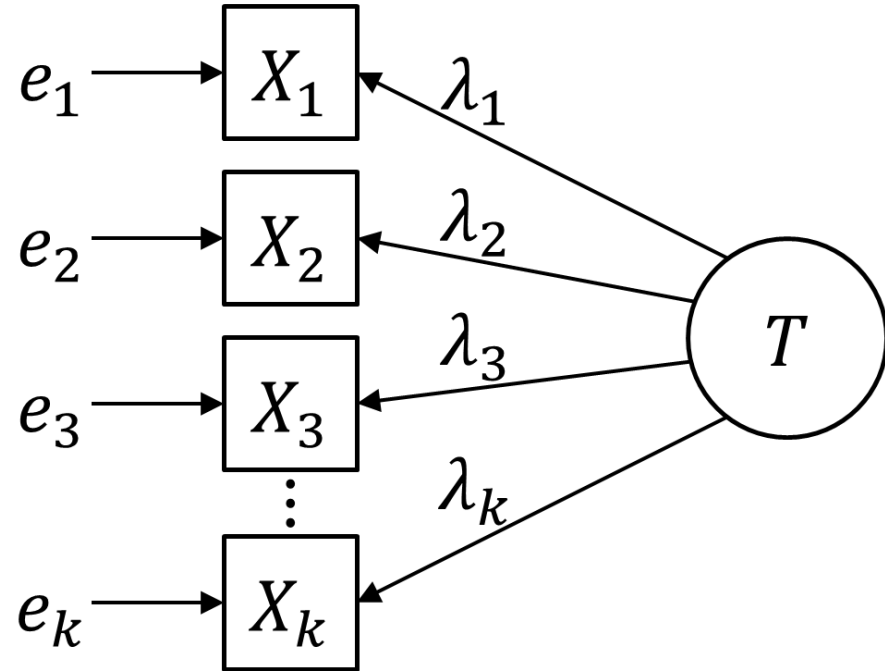
Example of a latent variable: Humor

- **Humor**: the quality of being amusing or comic, especially as expressed in literature or speech.

"Explaining humor is a lot like dissecting a frog: you learn a lot in the process but in the end you kill it."

-(attributed to) Mark Twain (by the internet)

What are some characteristics of humor that we can measure directly?

- Timing
- Content
- Structure (e.g., "knock-knock" joke vs. pun)
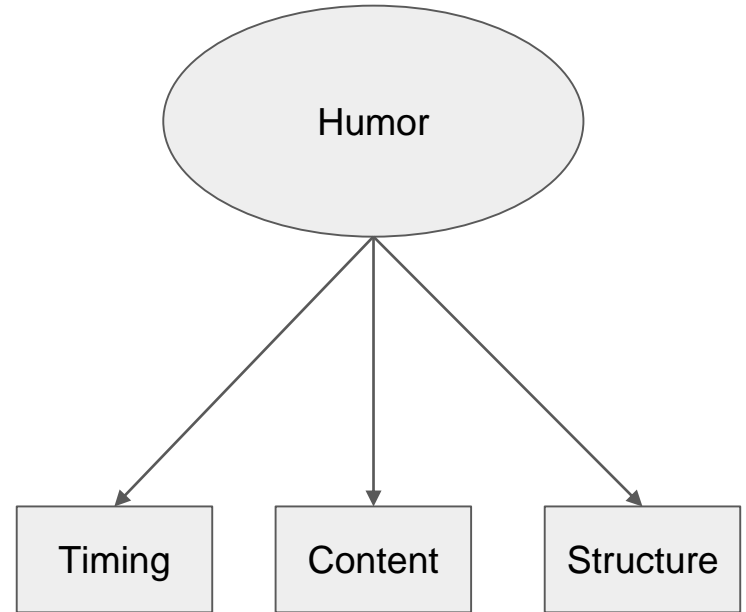
What is the value of latent variables?

- Allows us to **explicitly model** complex constructs in a manner that reflects that they are compositional (that is, that there is not one single measure/variable that adequately indexes that construct)
  - Depression
  - Frailty
  - Stress
  - Clinical net benefit

- They are a way to **account for the measurement error** inherent in constructs for which there is no "gold standard" for assessment.

What is the value of latent variables?

- They are **flexible** (i.e., they can be continuous (factors) or categorical (classes)), and can be estimated from many different types of data using most software packages.

- Allow us to **empirically** identify distinct groups (in the case of latent class models) instead of relying on **"artificial"** (e.g., diagnostic criteria that may not reflect sociocultural differences) or **convenient** (e.g., above/below the median on some measure) **thresholds** of identifying groups.

Extensions: Structural equation modeling (SEM)

You can **model the relationship among observed variables** (e.g., age, sex, race) **and latent variables** (i.e., depression, frailty), which themselves can have shared indicators
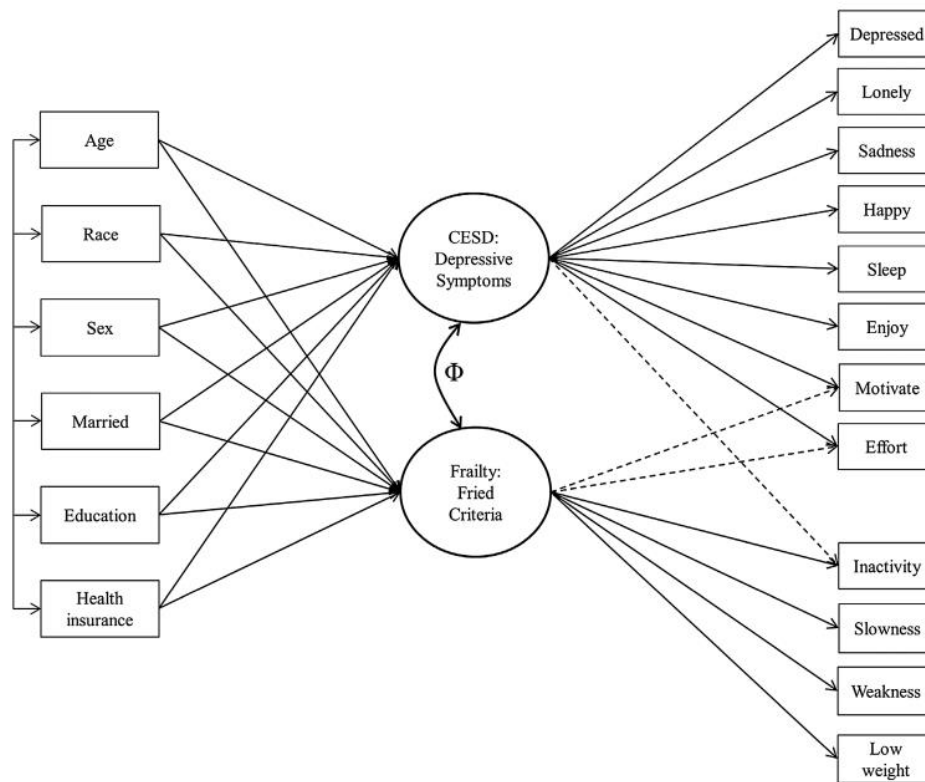

Figure 1. Heuristic example of correlated two-factor model adjusting for covariate influence. *Note.* Similar models were fit for each of three frailty definitions. Not all covariates included in analyses are depicted in figure. Dashed lines represent symptoms shared by frailty and depression.

Even more extensions

You can **compare different measurement models derived from theoretical considerations** (i.e., are depression and frailty just correlated, or does their correlation reflect an additional, second-order "vulnerability" factor?) using fit statistics (RMSEA, BIC, AIC, etc.)
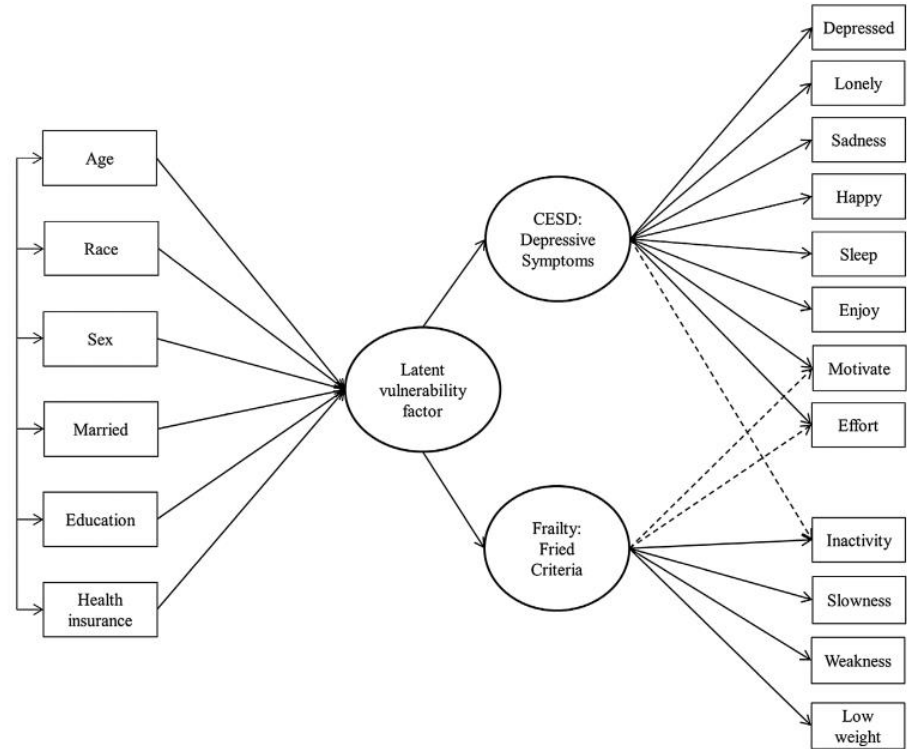

Figure 2. Example of second-order latent factor model. *Note.* Similar models were fit for each of three frailty definitions. Not all covariates included in analyses are depicted in figure. Dashed lines represent symptoms shared by frailty and depression.

Even more extensions

You can use latent variables to **characterize changes over time in observed variables** (e.g., to extract trajectories) using "growth" modeling…

…which you can then use to predict other outcomes (e.g., likelihood of developing frailty).

Class 1: Weight gain (1.4%) ⋯⋯ Class 2: Weight loss (1.7%) ─ ─
Class 3: Consistent overweight (90.1%) ── Class 4: Consistent obesity (6.8%) ──

Why would you not use latent variables in your research?

- Latent variables require assumptions that are often **not directly testable** (e.g., is your construct *really* continuous or could it be categorical? Do you have all the indicators of said construct?).
- Extracting some features of latent variables (e.g., number of factors, number of classes) requires **substantial input from the investigator**, which means that different investigators may interpret the same output differently.
- There are **other ways to represent complex constructs** (e.g., network analysis) that don't require you to assume that latent variables exist at all!

Further reading

- [Latent variables in psychology and the social sciences](#)