# ADVANCED DATA ANALYSIS FOR PSYCHOLOGICAL SCIENCE

## Introduction and general course information

**Luca Menghini Ph.D.**

luca.menghini@unipd.it

***

Master degree in Developmental and Educational Psychology

University of Padova

2023-2024

# The professor

Luca Menghini Ph.D.

Work & organizational psychologist

Postdoctoral research fellow in Applied psychology

& Quantitative research methods @uniTN

Professor on contract @uniPD

# My path

- 2014: Bsc in Work & Social Psych Sciences @uniPD

  *Biofeedback training for work stress management*

- 2016: Msc in Social, Work, & Communication Psych @uniPD

  *Psychophysiological workplace stress assessment protocol*

- 2017: Psychology consultance internship @InsidePerformance

  *Stress managament & biofedback training in professional sport drivers*

  + Research internship @uniPD *Actigraphy in sleep research* ← R 😉

- 2020: Ph.D. in Psychological Sciences @uniPD

  *Ecological momentary assessment of workplace stress* ← ΣΨ 😉

- 2020: Visiting scholar @SRI International (CA, USA)

  *Wearable sleep trackers, sleep and stress in adolescent insomnia*

- 2021: Postdoc @uniBO

  *Workaholism and daily fluctuations in blood pressure, emotional exhaustion, and sleep quality*

- 2022: Postdoc @uniTN *Youth between transitions, challenges, and opportunities*

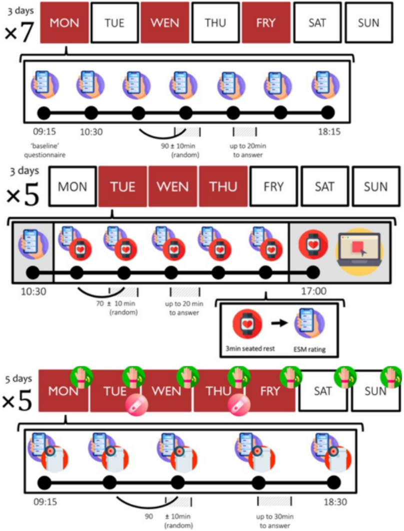
# Some of my studies related to the course content

- **Multilevel**: Menghini, L., Yüksel D., Baker, F. C., King, C., de Zambotti, M. (2023). Wearable and mobile technology to characterize daily patterns of sleep, stress, pre-sleep worry, and mood in adolescent insomnia. *Sleep Health 9*(1), 108-116. https://doi.org/10.1016/j.sleh.2022.11.006 [FULL-TEXT] [R CODE]

- **Multivariate**: Menghini, L., Balducci, C., & Toderi, S. (2022). Italian adaptation of the Warr's Job-related Affective Wellbeing Scale: Factorial structure and relationships with the HSE Management Standards Indicator Tool. *TPM – Testing, Psychometrics, Methodology in Applied Psychology, 29*(3), 309-325. https://doi.org/10.4473/TPM29.3.3 [FULL-TEXT] [R CODE]

- **Multilevel & Multivariate**: Menghini, L., Pastore, M., Balducci C. (2022). Workplace Stress in Real Time: Three Parsimonious Scales for the Experience Sampling Measurement of Stressors and Strain at Work. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000725 [FULL-TEXT] [R CODE]

# Intensive longitudinal designs



# Warnings :)

- I'm not a statistician
- I'm not a mathematical psychologist
- I'm not a programmer

I'm an Applied psychologist passionate about modeling and psychometrics.

Plus, this is my first time with this course: suggestions and critiques are welcomed!

Ethical code for psychology research: Explicit aknowledgement of limitations

# Contact & office hours

**Contact**: Moodle or mail to: luca.menghini@unipd.it

**Office hours**: TO DO

**Where**: Psico 1 pink building, ground floor, between the computer rooms

We can also schedule Zoom meetings

# Advanced data analysis for psychological science: Course overview

This course aims at providing basic notions of **multi-level & multi-variate** linear regression modeling, focusing on applications in developmental, educational, and applied psychology.

The course aims at transmitting **basic knowledge** on *linear mixed-effects regression* (LMER) and two common examples of multivariate techniques within the structural equation modeling (SEM) framework, namely *path analysis* & *confirmatory factor analysis* (CFA).

The course also aims at providing **practical competences** on advanced data analysis, with a particular emphasis on data preparation and pre-processing, model fit, evaluation, and selection criteria, coefficient interpretation, and data visualization.

The course is characterized by an **applied approach** that prioritizes real case studies and includes practical exercises using R.

# Prerequisites

Students should have good knowledge about basic concepts linked to probability theory and associated topics (e.g., random variables, probability distributions, hypothesis testing), including **linear regression modeling**.

# Course contents

1. Intro and course info 📍

**Multi-level**

2. Introduction: From `lm()` to `lmer()`
3. Data preparation R
4. Model fit & random effects R
5. Model evaluation & selection R
6. Coefficient interpretation R
7. Generalized models `glmer()`, Bayesian LMER, & power analysis 🔬

**Multi-variate**

8. Introduction: From `lm()` to `sem()`
9. Observed variables & path analysis
10. Data preparation R
11. Model evaluation & selection R
12. Coefficient interpretation R
13. Latent variables & CFA
14. Full SEM pipeline R
15. Multilevel SEM, Mediation, Bayesian SEM, power analysis 🔬

R = Practical exercise sessions with R (bring your PC!)

🔬 = In-depth topics (not for the exam!) ← slides with blue boxes and the microscope icon

# When & where

The course will last 42 hours (6 ECTS).

All lectures will be delivered in the Psico 2 gray building, room 3F - via Venezia 12.

| Day | Date | Time | Room |
|---|---|---|---|
| 1 | 10-4 (wed) | 12:30-14:30 | 3F |
| 2 | 10-5 (thu) | 08:30-10:30 | 3F |
| 3 | 10-11 (wed) | 12:30-14:30 | 3F |
| 4 | 10-12 (thu) | 08:30-10:30 | 3F |
| 5 | 10-18 (wed) | 12:30-14:30 | 3F |
| 6 | 10-19 (thu) | 08:30-10:30 | 3F |
| 7 | 10-25 (wed) | 12:30-14:30 | 3F |
| 8 | 10-26 (thu) | 08:30-10:30 | 3F |
| 9 | 11-1 (wed) | 12:30-14:30 | 3F |
| 10 | 11-2 (thu) | 08:30-10:30 | 3F |
| 11 | 11-8 (wed) | 12:30-14:30 | 3F |

| Day | Date | Time | Room |
|---|---|---|---|
| 12 | 11-9 (thu) | 08:30-10:30 | 3F |
| 13 | 11-15 (wed) | 12:30-14:30 | 3F |
| 14 | 11-16 (thu) | 08:30-10:30 | 3F |
| 15 | 11-22 (wed) | 12:30-14:30 | 3F |
| 16 | 11-23 (thu) | 08:30-10:30 | 3F |
| 17 | 11-29 (wed) | 12:30-14:30 | 3F |
| 18 | 11-30 (thu) | 08:30-10:30 | 3F |
| 19 | 12-6 (wed) | 12:30-14:30 | 3F |
| 20 | 12-7 (thu) | 08:30-10:30 | 3F |
| 21 | 12-13 (wed) | 12:30-14:30 | 3F |

# Course materials

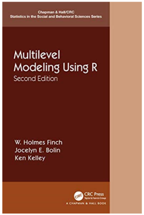
All course materials can be accessed from the Moodle page of the course ***and*** from https://github.com/Luca-Menghini/advancedDataAnalysis-course

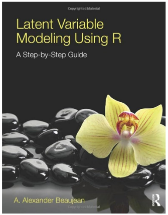
The contents required by the exam are exhaustively covered in the **main course slides**:

1. **Intro & course info** (the present slides)
2. **Multilevel modeling**
3. **Multivariate modeling**

***Suggested* textbooks to deepen the topics of the course:**

- Finch, W. H., Bolin, J. E., Kelley, K., Multilevel Modeling Using R (2nd edition). Boca Raton: CRC Press, 2014
- Beaujean, A. A., Latent Variable Modeling Using R. A Step-by-Step Guide. New York: Routledge, 2014





# Course slides

The course slides are structured by intermixing theory, R code, plots, examples, and exercises.

The R code used in any exercise/example is also provided.

Slides with blue boxes and the microscope 🔬 icon cover in-depth but still useful topics that will be possibly presented but are not part of the core course topics and related exam!

All course materials can be accessed from Moodle
***and*** from https://github.com/Luca-Menghini/advancedDataAnalysis-course

# Additional resources

Additional resources that are not presented during classes will be also available from Moodle. These will include published papers and online resources, R code and exercises, extra slides, and other.

For instance, you can already find the `R-intro.pdf` extra slides (introduction to R), and you can already give a look at the “Latent Variable Modeling using R” book website (e.g., “R syntax” section): https://blogs.baylor.edu/rlatentvariable/

**PSICOSTAT meetings & workshops**:

Interdisciplinary research group on quantitative psychology, psychometrics, psychological testing, & statistics - monthly online meetings + weekly in person workshops
https://psicostat.dpss.psy.unipd.it/index.html

# Teaching modalities

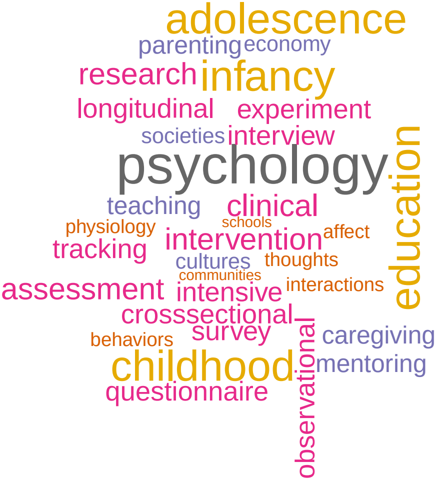
Frontal theoretical sessions on the rationale of the analytical techniques focused by the course

Practical sessions with individual and group exercises

Practical sessions will be based on the freely-available R software.
Students are encouraged to bring their **laptops**, if possible.

The course will emphasize practical examples and **cases studies**
in developmental, educational, and applied psychology.

# Case studies:
# What are your research and/or applied interests?



# Attending the course: “PACATE”

**Participation**: You are expected to contribute to the class by participating in class discussion and working with each other during practical sessions. If you find something unclear or discordant with other information, please tell it to the professor. If you find it uncomfortable to speak up in class, feel free to contact the professor and work on this skill.

**Attendance**: Class attendance is not mandatory but encouraged. It is recommended to gradually but constantly familiarize with the content of the course.

**Collaboration**: Please, help each other, for instance, by working together on practical sessions and assignments, and/or by exchanging notes and useful materials for the exam.

**Assignments**: Over the course, several exercises/homework will be *suggested* to consolidate the course contents. While some of these will be discussed in class, feel free to contact the professor if you find any issue with the assignments.

**Timekeeping**: You are expected to be on time. You should be in your seat and ready to begin class when the class starts.

**ExeRcise!**

# An inseparable companion

`R` is a programming language and a programming environment for **statistical computing** and **graphics**.

It is based on the `S` language (Becker & Chambers, 1984), subsequently used to develop the `S-Plus` sofware and then `R`, originally created in 1996 by Ross Ihaka and Robert Gentleman.

Today, it is supported by an international research group (R Core Team and R Foundation for Statistical Computing) that periodically update (each year) the base sofware (***Base R***).

Progressive and exponential inclusion of new **packages** that extend its capabilities.

https://www.r-project.org/

# A bottom-up resource

R provides a wide range of statistical and graphical techniques. It is designed to be **user-friendly** but at the same time to generate **high-quality outputs** (graphics, tables, and reports with equations, mathematical symbols, etc.).

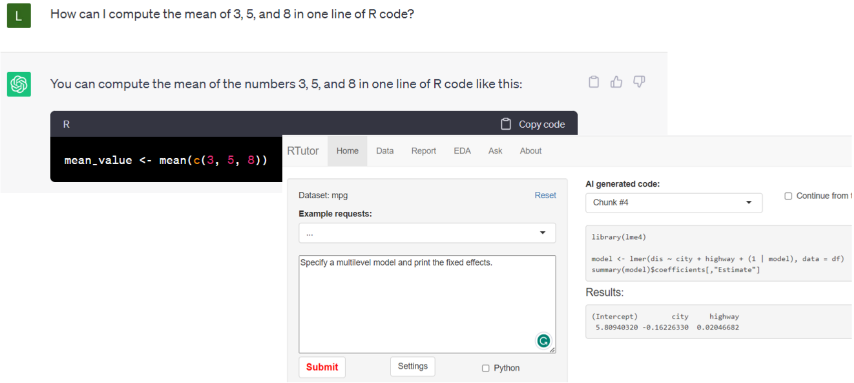
Optimized default functions
+ dedicated packages
+ possibility to fully control.

Differently from other statistical software that implement multilevel and multivariate analyses (e.g., Mplus), R is a **free sofware** (GNU General Public Licence) that can be used anywhere worldwide, it is an **open-source software** (all functions are documented and can be inspected in detail), and **works on all main OS**: Windows, MacOS, and UNIX (e.g. Linux)

Moreover, there is a massive community of useRs: For any issue, just Google it! ☻

# Even better than googling

Try https://chat.openai.com/ or https://rtutor.ai/



L How can I compute the mean of 3, 5, and 8 in one line of R code?

You can compute the mean of the numbers 3, 5, and 8 in one line of R code like this:

```r
mean_value <- mean(c(3, 5, 8))
```

Dataset: mpg

Example requests: ...

Specify a multilevel model and print the fixed effects.

□ Python

AI generated code: Chunk #4

□ Continue from t

```r
library(lme4)

model <- lmer(dis ~ city + highway + (1 | model), data = df)
summary(model)$coefficients[,"Estimate"]
```

Results:

```
(Intercept)        city     highway 
 5.80940320 -0.16226330  0.02046682 
```

# Introduction to R: Additional materials

**Extra slides on Moodle/GitHub**

- `extra/R-intro.pdf`: How to install and get started with `R` and `RStudio`, elementary commands, `R` objects, functions, and workspace, how to read and export datasets, `R` graphics, and linear models
- `extra/ggplot2-intro.pdf`: Introduction to the `ggplot2` package for advanced graphics

**Free tutorials**

- Navarro, D. Learning statistics with R: A tutorial for psychology students and other beginners - https://learningstatisticswithr.com/
- `learnr`: an R package for learning how to use R https://rstudio.github.io/learnr/
- excellent STAT545: Data wrangling, exploration, and analysis with R https://stat545.com/

# Key R packages used in the course

The course uses several packages with customized and optimized functions. Here are the main packages used in the course (and the code to install all of them):

```r
pckg <- c("lme4","lavaan",
          "plyr","reshape2","sjPlot")
install.packages(pckg)
```

Note: This course does not use `tidyverse` packages and syntax (see https://www.tidyverse.org/ ), but relies on R Base.

# Some key R functions used in the course

Aggregating scores by group

```
aggregate(x = sleep$extra,
          by = list(sleep$group), FUN = mean)
```

```
##   Group.1    x
## 1       1 0.75
## 2       2 2.33
```

Merging wide- and long-form datasets

```
new <- join(long, wide, by = "ID", type = "left")
```

Fitting LMER models and printing fixed effects

```
fit <- lmer(extra ~ group + (1|ID), data = sleep)
fixef(fit)
```

```
## (Intercept)      group2
##        0.75        1.58
```

Fitting SEM and printing coefficients

```
fit <- sem("visual =~ x1 + x2 + x3
           textual =~ x4 + x5 + x6
           visual ~ textual",
           data=HolzingerSwineford1939)
```

```
standardizedsolution(fit)[1:7,1:4]
```

| lhs | op | rhs | est.std |
|---|---|---|---|
| visual | =~ | x1 | 0.78 |
| visual | =~ | x2 | 0.43 |
| visual | =~ | x3 | 0.57 |
| textual | =~ | x4 | 0.85 |
| textual | =~ | x5 | 0.85 |
| textual | =~ | x6 | 0.84 |
| visual | ~ | textual | 0.46 |

# When & where

All exam sessions will take place in the Psico 2 gray building, room 3L
Via Venezia 12, Padova - 35131

| Session | Date | Time | Room |
|---|---|---|---|
| 1. Jan | 2024-01-17 | 14:30 | 3L |
| 2. Feb | 2024-02-14 | 14:30 | 3L |
| 3. Jun | 2024-06-10 | 14:30 | 3L |
| 4. Jul | 2024-07-10 | 14:30 | 3L |
| 5. Aug | 2024-08-10 | 14:30 | 3L |

# Exam structure & contents

The final exam will be **written** and will last **40 minutes**.
The exam will consist of **31 closed-ended questions** on:

- theoretical topics covered by the course
- data analysis exercises using R (analysis of case studies) based on the procedures learned during the course.

The contents required by the exam are exhaustively covered in the main course slides. The exam score will be computed as the sum of the scores obtained to the 31 questions.

More information about the final exam will be provided later, along with an exam simulation.

# Example questions

- theoretical topics: to do
- data analysis exercise: to do

# Multi - LEVEL & Multi - VARIATE

Advanced statistical techniques to deal with **large and complex data structures**

## Multilevel regression model

To be used with **hierarchical data structures** where lower-level observations (statistical units) are ***nested*** within higher-level variables (clusters).

*Linear mixed-effects regression* (LMER) allows to estimate **fixed effects** that are constant across all clusters + **random effects** varying from cluster to cluster.

LMER is widely applied in developmental and educational psychology:

- students → classes → schools
- experiences → days → individuals
- trials → items & individuals

## Multivariate regression model

To be used to account for the multivariate reality of psychosocial phenomena, where **multiple variables interact** at the same time (e.g., multiple outcomes, mediation).

*Path model* = Pictorial representation (diagram) of a theory of variable relationships (*structural model*)

*Latent variable model* = representation of the relationships that form the ***latent variables*** used in a structural model

*Structural equation model* (SEM) = full model composed by the latent and structural part

# Linear models as the common root

**Regression models** aim to establish whether two variables are in a asymmetric functional relationship, and particularly to quantify the extent to which one `X` variable (*independent* or *predictor*) influences the `Y` variable (*dependent* or *response*)

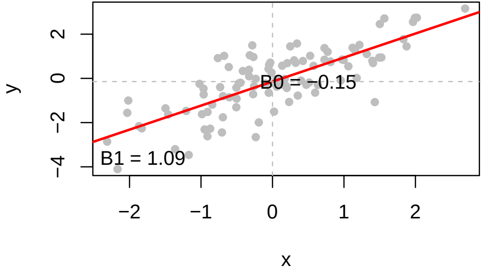
**Linear regression** allows to determinate the link between two variables as expressed by a linear function: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Such a function can be graphically represented as a **straight line** where

$\beta_0$ is the **intercept** (value assumed by `Y` when `X` $= 0$)

$\beta_1$ is the **slope coefficient** (predicted change in `Y` when `X` increases by 1 unit)

$\epsilon$ is the **residual variance** (distance from the regression line)



# The only three formulas to keep in mind

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Mixed-effects model:

$$Y_{ij} = (\beta_0 + \lambda_{0j}) + (\beta_1 + \lambda_{1j}) X_{ij} + \epsilon_{ij}$$

For each observation $i$ and each cluster $j$, the intercept and the slope are decomposed into the **fixed** components $\beta_0$ and $\beta_1$ referred to the whole sample, and the **random** components $\lambda_{0j}$ and $\lambda_{1j}$ randomly varying between clusters

Structural equation model:

$$\begin{cases} x_i = \Lambda_x \xi_i + \delta_i \ (meas.\ x) \\ y_i = \Lambda_y \eta_i + \epsilon_i \ (meas.\ y) \\ \eta_i = \Gamma \xi_i + \zeta_i \ (struct.) \end{cases}$$

For each observation $i$, the **measurement model** (first two lines) clarifies the relationships $\Lambda$ between the *observed variables* $x$ and $y$ and the corresponding *latent variables* $\xi$ and $\eta$, whereas the **structural model** (third line) clarifies the relationship $\Gamma$ between the two latent variables.

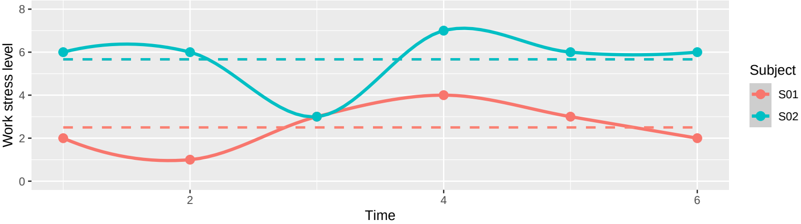
# Multilevel models: Let’s make the visuals talk

**Visual introduction to multilevel modeling**:

http://mfviz.com/hierarchical-models

**Multilevel modeling in repeated measures & longitudinal designs**

When a random variable is measured repeatedly over time from different individuals, observations are *nested* within individuals and multilevel modeling can be used to **partition the variance** into the ***within-subject*** (level 1) and ***between-subjects*** (level 2) components.



# Multilevel models: Fixed vs. Random effects

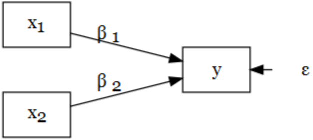
In the literature, multilevel modeling is sometimes called with different terms, e.g., *hierarchical linear modeling*, *random slope models*, *variance component models*, . . .

All these models are part of the broader ***mixed-effects models*** family, identifying models with both fixed and random effects:

- **Fixed effects**: effects that remains constant across all clusters, whose *levels* are exhaustively considered (e.g., gender, levels of a Likert scale) and generally controlled by the researcher (e.g., experimental conditions)
- **Random effects**: effects that vary from cluster to cluster, whose *levels* are randomly sampled from a population (e.g., schools, participants, days, experimental stimuli)

# Multivariate models: Let's make the visuals talk

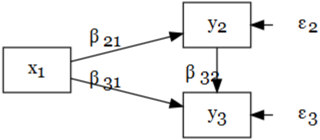
**Linear regression**: determining the link between a dependent and an independent variables through linear functions like: $y = \beta_1 x_1 + \beta_2 X_2 + \epsilon$



As a limitation, linear models can only predict **one dependent variable at time** with a single equation. They can be *univariate* (without predictors) or *bivariate* (with predictors).

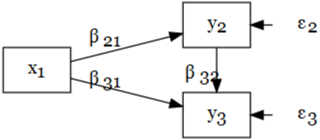
**Structural equation models** (SEM) are *multivariate* models that allow simultaneously modeling multiple ~~dependent~~ *endogenous* variables with a **system of equations**:

$$\begin{cases} y_2 = \beta_{21} x_1 + \epsilon_2 \\ \\ y_3 = \beta_{31} x_1 + \beta_{32} Y_2 + \epsilon_3 \end{cases}$$



# Multivariate models: exogenous vs. endogenous

$$
\begin{cases}
y_2 = \beta_{21} x_1 + \epsilon_2 \\
y_3 = \beta_{31} x_1 + \beta_{32} y_2 + \epsilon_3
\end{cases}
$$

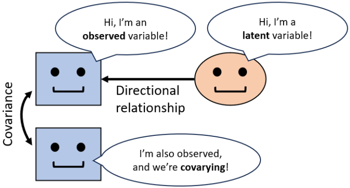


In SEM, the classic independent vs. dependent classification is replaced with a more meaningful one:

- **Exogenous variables** ($X_1$): without a direct ‘cause’ from inside the model (*predictors*), without error estimate
- **Endogenous variables** ($Y_2$, $Y_3$): directly ‘caused’ from inside the model (*predictors* & *outcomes*), with error estimate $\epsilon$

# Multivariate models: observed vs. latent

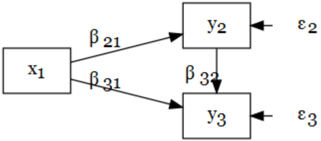
A further advantage of SEM is to distinguish between observed vs. latent variables


Hi, I'm an **observed** variable!
Hi, I'm a **latent** variable!
Directional relationship
Covariance
I'm also observed, and we're **covarying**!

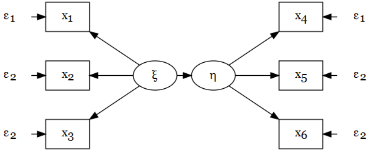
**Observed variables**: directly observable and measurable (e.g., heart rate), represented by *squares* and the *lowercase letters* $x$ (exogenous) and $y$ (endogenous)

**Latent variables**: hypothetical and not directly measurable but **indexed** by one or multiple observed variables (e.g., happiness), represented by *circles* and the *greek letters* $\xi$ (exogenous) and $\eta$ (endogenous)

When including **observed variables only**, SEM are called **path analysis**, which is widely used to model complex multivariate relationships (e.g., *mediation models*):



When **both observed and latent** variables are included, we can talk of ‘full SEM’:



# Multilevel & multivariate models: It's a matter of theory!

While *any model is a formal representation of a theory* (Bollen, 1989), the formulation of a multilevel and/or multivariate model is particularly dependent on the underlying theoretical model.

**Multilevel modeling**:

Theories determinate whether a clustering variable is meaningful or not, the number of levels (e.g., individuals, days, weeks, schools) to be considered, and whether a given construct can be meaningfully attributed to a given level (e.g., happy people, happy days, happy weeks, happy schools).

**SEM**:

Theories determinate both how a latent variable is reflected by a set of observed variables (**measurement model**) and what are the regression-like relationships among the variables (**structural model** ~ *path analysis*).

Note: when a SEM is analyzed without a structural model, it is usually called **confirmatory factor analysis** (CFA).

# To be continued. . .

- Any question?
- Next lecture: LM recap or LMER?
- Recap on how to use R?

# Credits

The present slides are partially based on:

- Altoè, G. (2023) Corso Modelli lineari generalizzati ad effetti misti - 2023. https://osf.io/b7tkp/
- Beaujean, A. A. (2014) Latent Variable Modeling Using R. A Step-by-Step Guide. New York: Routledge
- Finch, W. H., Bolin, J. E., Kelley, K. (2014). Multilevel Modeling Using R (2nd edition). Boca Raton: CRC Press
- Pastore, M. (2015). Analisi dei dati in psicologie (e applicazioni in R). Il Mulino.

# Useful resources: Multilevel

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412.
- Bliese, P. (2022). Multilevel modeling in R (2.7).
  https://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf
- McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.
- Pinheiro, J., & Bates, D. (2006). Mixed-effects models in S and S-PLUS. Springer science & business media.
  see also Bates, D. (2022). lme4: Mixed-effects modeling with R.
  https://stat.ethz.ch/~maechler/MEMo-pages/lMMwR.pdf

# Useful resources: Multivariate

- Kline, R.B. (2005). Principles and Practice of Structural Equation Modeling. Guilford Press, NY.
- Lin, J. Introduction to structural equation modeling (SEM) in R with lavaan. https://stats.oarc.ucla.edu/r/seminars/rsem/
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*, 1-36.