lavaan: past, present and future

Yves Rosseel
Department of Data Analysis
Ghent University – Belgium

Meeting of the SEM Working Group
Warsaw, 16–17 April, 2026

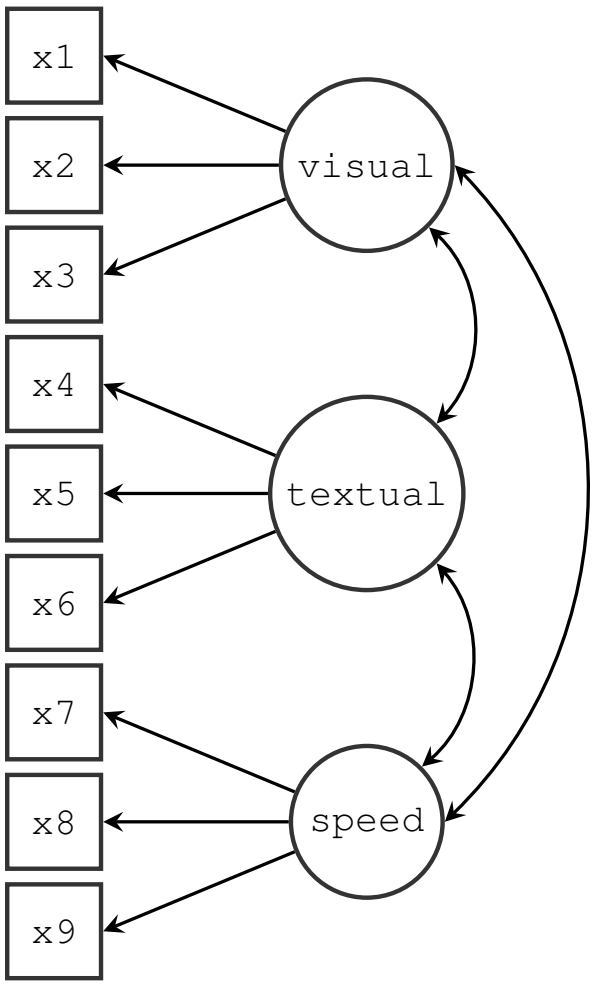
the past (2000–2008)

- research: mathematical psychology, neuro-imaging data analysis (no SEM)
- statistical consultancy:
  - I often used LISREL, EQS or Mplus, depending on the client
  - mostly just confirmatory factor analysis (CFA); often very repetitive (same model, multiple datasets)
  - around 2008, the only option in R was the **sem** package, which was too limited for my purposes
- teaching:
  - I was teaching a data analysis course that included a chapter on CFA, and we were using R for everything else
- the initial plan:
  - create a small (non-public) R package to do only 1 thing: CFA
  - first package (March 2009, never published): **cfa2000**

first package: 'cfa2000', March 2009

```
Package: cfa2000
Type: Package
Title: Confirmatory Factor Analysis
Version: 0.9-3
Date: 2009-03-27
Author: Yves Rosseel <yves.rosseel@ugent.be>, with contributions from Jan Lammert
Maintainer: Jan Lammertyn <jan.lammertyn@ugent.be>
Description: Confirmatory Factor Analysis, better and faster.
Depends: methods
License: GPL version 2 or later
LazyLoad: yes
LazyData: yes
Packaged: Tue Mar 24 13:48:59 2009; michael
```

cfa2000 example: Holzinger & Swineford (1939) 3-factor CFA

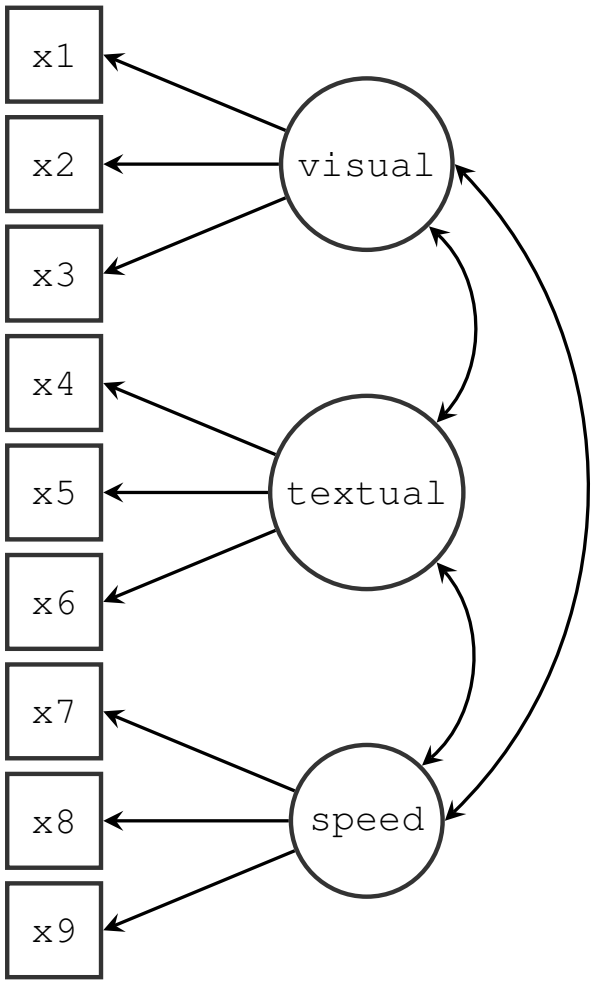



```
library(cfa2000)

# specify 3-factor CFA model
measurement.model <-
    list( visual  = c("x1","x2","x3"),
          textual = c("x4","x5","x6"),
          speed   = c("x7","x8","x9") )

# fit the model
fit <- cfa(measurement.model = measurement.model,
           data = HolzingerSwineford1939)
summary(fit)
```

cfa2000 partial output

```
Model converged normally after 35 iterations (0.146s)

  Chi-square test full model          85.306
  Degrees of freedom                     24
  P-value                            0.0000

Factor loadings:     Estimate      S.E.   z value  Pr(>|z|)
  visual
     x1                 1.000
     x2                 0.554     0.100     5.554     0.000
     x3                 0.729     0.109     6.685     0.000
  textual
     x4                 1.000
     x5                 1.113     0.065    17.014     0.000
     x6                 0.926     0.055    16.703     0.000
  speed
     x7                 1.000
     x8                 1.180     0.165     7.152     0.000
     x9                 1.082     0.151     7.155     0.000

Factor var/cov:      Estimate      S.E.   z value  Pr(>|z|)
  visual
     visual             0.812     0.146     5.564     0.000
     textual            0.410     0.074     5.552     0.000
     speed              0.263     0.056     4.660     0.000
...
```

cfa2000, August 2009, using formula-like expressions



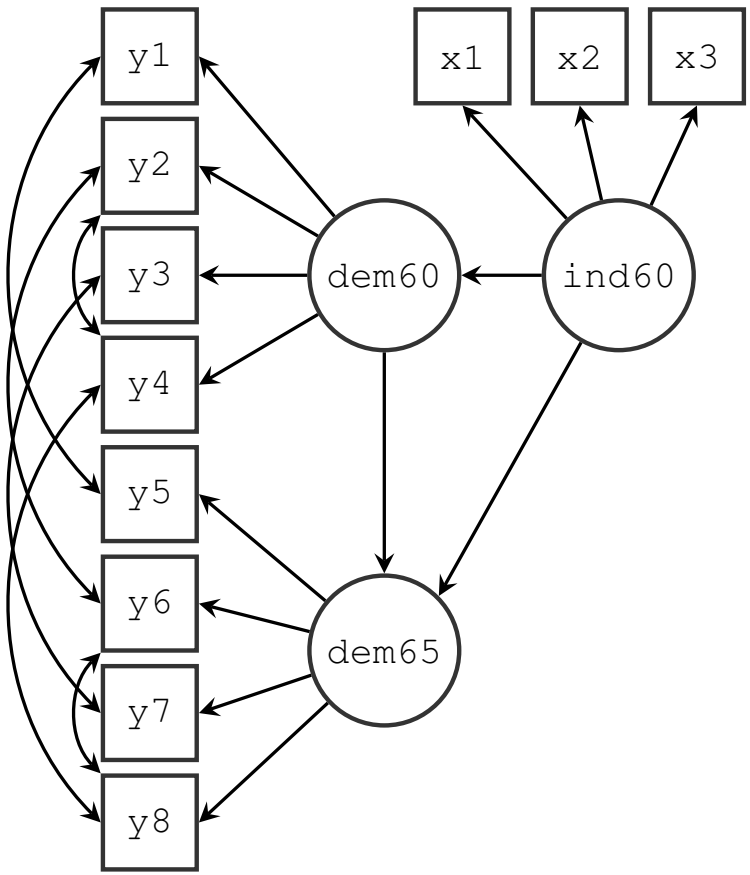

```
visual  =~ x1 + x2 + x3
textual =~ x4 + x5 + x6
speed   =~ x7 + x8 + x9

fit <- cfa(measurement.model = list(visual,
                                    textual,
                                    speed),
           data = HolzingerSwineford1939)

summary(fit)


fit.measures(fit, c("cfi", "rmsea", "srmr"))
```

- nice and easy
- but what about exogenous covariates?
- we need a SEM package after all

second package: ‘semplus’, September 2009

```
Package: semplus
Type: Package
Title: Structural Equation Modeling
Version: 0.9-9
Date: 2009-09-16
Author: Yves Rosseel <yves.rosseel@ugent.be>
Maintainer: Yves Rosseel <yves.rosseel@ugent.be>
Description: Structural Equation Modeling with a formula-based interface
Depends: methods, MASS
License: GPL version 2 or later
LazyLoad: yes
LazyData: yes
Packaged: 2009-10-13 08:18:48 UTC; yves
```

semplus: political democracy example




```
# measurement part
mm <- list(ind60 =~ x1 + x2 + x3,
           dem60 =~ y1 + y2 + y3 + y4,
           dem65 =~ y5 + y6 + y7 + y8)

# correlated errors
ce <- list(y1 ~~ y5,
           y2 ~~ y4 + y6,
           y3 ~~ y7,
           y4 ~~ y8,
           y6 ~~ y8)

# structural part
eqs <- list(dem60 ~ ind60,
            dem65 ~ ind60 + dem60)

fit <- sem(measurement.model = mm,
           eqs = eqs,
           ce = ce,
           data = BollenDemocracy)
```

Jan 2010 – semplus using list() to specify the model

```
model <- list(

    # latent variable definitions
        ind60 =~ x1 + x2 + x3,
        dem60 =~ y1 + y2 + y3 + y4,
        dem65 =~ y5 + y6 + y7 + y8,

    # regressions
        dem60 ~ ind60,
        dem65 ~ ind60 + dem60,

    # residual (co)variances
        y1 ~~ y5,
        y2 ~~ y4 + y6,
        y3 ~~ y7,
        y4 ~~ y8,
        y6 ~~ y8
)
```

March 2010 – semplus using a string literal

```
model <- '
  # latent variable definitions
      ind60 =~ x1 + x2 + x3
      dem60 =~ y1 + y2 + y3 + y4
      dem65 =~ y5 + y6 + y7 + y8

  # regressions
      dem60 ~ ind60
      dem65 ~ ind60 + dem60

  # residual correlations
      y1 ~~ y5
      y2 ~~ y4 + y6
      y3 ~~ y7
      y4 ~~ y8
      y6 ~~ y8
'

fit <- sem(model, data = PoliticalDemocracy)
```

from semplus to lavaan

- the package was named ‘semplus’ because it could do ‘more’ than the R package sem
- and it contained the word ‘mplus’
- I contacted the Mplus team (24-02-2010), with some technical questions
- and received an email back (03-03-2010) saying:

> *We own the Mplus trademark. Using the name “semplus” can be construed as a trademark infringement and might also imply our endorsement.*

- our legal department was eager to fight them
- eventually, I changed the name to ‘lavaan’ (latent variable analysis)
- (I never got an answer from Mplus for my technical questions)

lift off

- lavaan 0.3-1 (about 6470 lines) was released on CRAN on 11 May 2010
- presentation at useR 2010 (NIST, Gaithersburg, Maryland, USA)
  - no interest at all...
- invited speaker at Psychoco (February 2011) in Tübingen:
  - Achim Zeileis proposed to create a 'special volume' for the journal of statistical software on 'Psychometric Computing in R': https://www.jstatsoft.org/issue/view/v048
  - this is where the 'lavaan paper' was published (in 2012)
  - the same issue also contains the 'qgraph paper' from Sacha Epskamp
- Todd Little invited me to the University of Kansas (March 2012); this is where I met Terrence Jorgensen, and where the idea for 'semTools' was born
- from 2013 onwards: many workshops and presentations

why do we need lavaan?

- original propaganda from 2010–2013 . . .
  1. **lavaan** is for statisticians working in the field of SEM (and beyond)
     - it seems unfortunate that new developments in this field are hindered by the lack of open source software that researchers can use to implement their latest ideas
  2. **lavaan** is for teachers
     - teaching these techniques to students was often complicated by having to choose one of the commercial packages
  3. **lavaan** is for applied researchers
     - keep the syntax simple, provide all the features they need
- . . . still true today

the next years

- more and more features were added
- HUGE step: 0.5 added support for categorical data (binary/ordinal)
  - getting the 'asymptotic covariance matrix of the sample statistics' right (for thresholds, polychoric correlations, slopes, variances, ...) was a challenge
- more attention for:
  - optimization, scaling, stopping criteria, ...
  - numerical stability, numerical methods
  - what to do if a covariance matrix is not positive-definite?
  - ...
- biggest challenge in the early years:
  - the lavaan output was not (always) identical to the output of other (commercial) packages

lavaan gives (slightly) different results!

- example: Satorra-Bentler scaled test statistic for a 3-factor CFA model using the 'classic' Holzinger and Swineford (1939) data (N=301)

program	SB test statistic
lavaan 0.5-22	80.872
Mplus 7.11	81.908
EQS 6.1	81.141
LISREL 8.72	77.396

- experts (often) could not explain these differences
- users switching to lavaan complained and believed that lavaan's results could not be trusted
- that is why we added the 'mimic' argument (e.g., `mimic = "Mplus"`)

design mistakes

- lavaan grew organically, with no (design) plan whatsoever
- initial version: only single group; big change to internal housekeeping if multiple groups are involved
- multilevel models: within/between levels, but not the same as multiple groups
- multigroup-multilevel: from groups to blocks
- same engine for every model
- matrix based (while scalar methods may scale better)
- not scalable (try to fit a model with P=1000 variables)
- S4 classes
- not modular (enough)
- ...

but also some good choices

- lavaan model syntax
- the parameter table (internal representation of the model)
- consistency in output
- text-based summary() output for users
- data-frame parameterEstimates() output for scripting
- backwards compatible (syntax from 2010 still runs)
- the nlminb() optimizer
- no C/C++ code (as a quick fix for speed)
- ...

lavaan today

- current version 0.6-21 (about 100,000 lines)
- the official website:

  https://lavaan.org

- the lavaan paper:

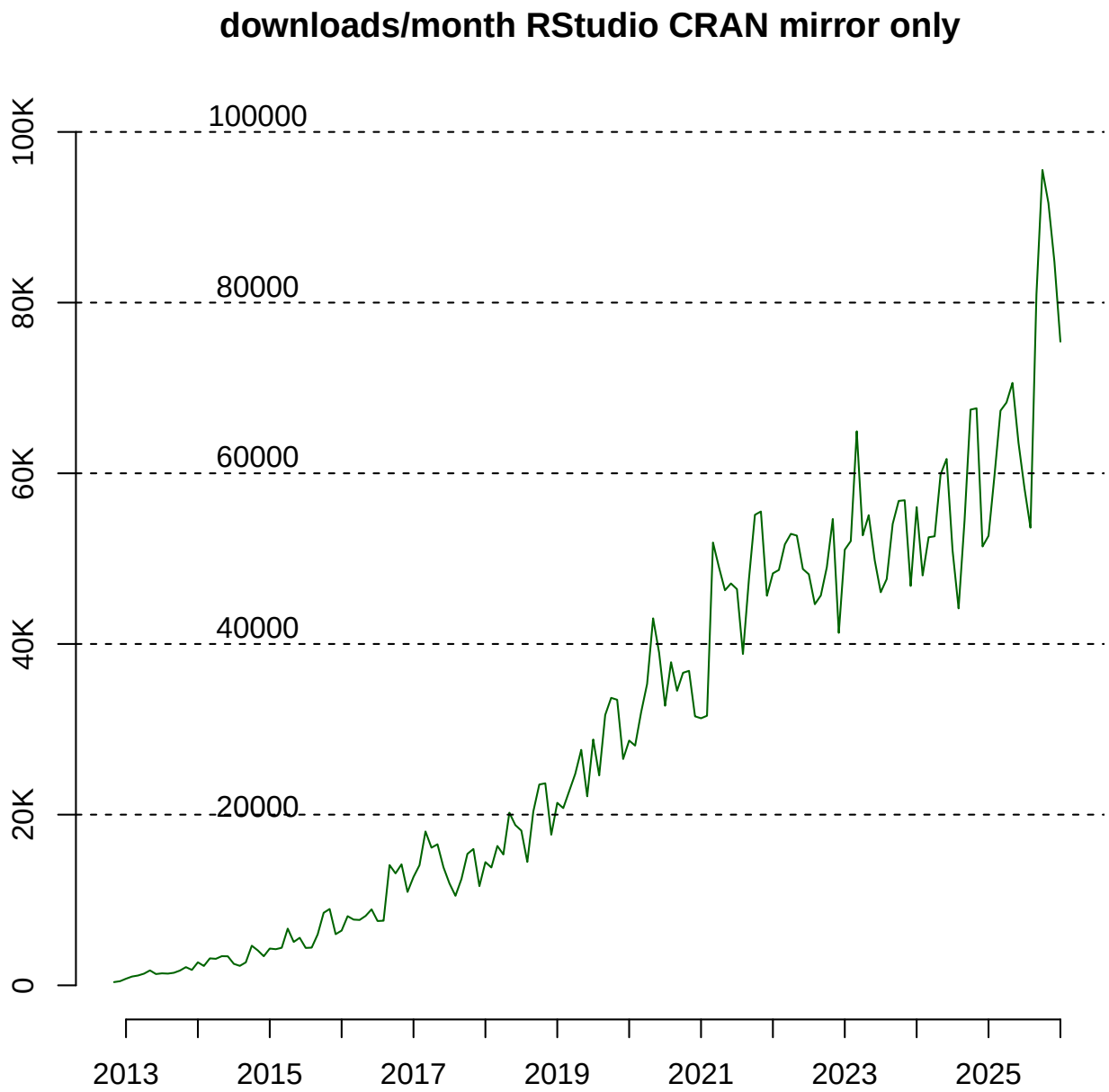
  Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

- lavaan source code:

  https://github.com/yrosseel/lavaan

- lavaan discussion group (mailing list)

  https://groups.google.com/d/forum/lavaan


downloads/month RStudio CRAN mirror only
100000
80000
60000
40000
20000
0
100K
80K
60K
40K
20K
2013
2015
2017
2019
2021
2023
2025

lavaan today (2)

- lavaan is now a mainstream package for SEM
- lavaan is widely used for both teaching and research
- the lavaan 'ecosystem' contains about 190 packages that depend on, or extend lavaan: AICcmodavg AlignLV aridagri autoFC bain Bayesrel bayestestR betaselectr bfw BifactorIndicesCalculator BIFIEsurvey blavaan bmem bmemLavaan bnpa bootnet broom broom.helpers bruceR calms ccpsyc CIEE coefficientalpha conmet CoTiMA covsim crosslag cSEM cvsem dagitty detectR discnorm dySEM EasyMx eatRep EFAtools EffectLiteR effectsize EGAnet EMAS EMgaussian esem ezCutoffs FAfA faoutlier FCO flps fSRM ggdiagram ggsem gimme gorica GRShiny ICED IIVpredictor influence.SEM INLAvaan insight IPV jmv JSmediation JWileymisc kfa konfound latcontrol latentFactoR lavaan.mi lavaan.printer lavaan.shiny lavaanExtra lavaangui lavaanPlot lavDiag lavinteract lcsm lessSEM LSAmitR lsl lslx LSMjml lvnet mantar manymome manymome.table MBESS MCCM measureR merDeriv metaSEM MIIVefa MIIVsem misty mitml mmibain modelbpp modelsummary modsem MonteCarloSEM MplusTrees multid multilevelTools negligible netCoin networksem nlsem NlsyLinks nonnest2 Omisc parameters paramtest pathmodelfit performance petersenlab phantSEM plavaan plssem pompom power4mome powerNLSEM powRICLPM primer profileR projectLSA PROsetta pscore psych psychonetrics psychTools pwr2ppl qgraph quest RAMpath regmed regsem reliacoef REMLA report restriktor RMediation rmedsem Rnest RSA RSAtools rsem RSP see semboottools SEMdeep semdrw semEffect semfindr SEMgraph semhelpinghands seminr semlbci semlrtp semmcci semnova semPlot semPower semptools SEMsens SEMsensitivity semTests semTools semtree ShortForm silp simsem simstandard siren sirt specr sphereML Spower STARTS stdmod stim stuart tabledown TAM ThurMod thurstonianIRT tidySEM tidystats ufs umx unusualprofile WebPower worcs wsMed

growing pains

- the more users, the larger my responsibility
- I spend more time 'testing' than coding
- each update is something of a nightmare
  - you shall not break a package that depends on your package! (CRAN)
  - you shall not alter the way the output looks!
  - you shall not change the numbers in the output! (same model, same data)
  - you shall not make any mistakes! (lavaan users have come to expect that everything works perfectly, all the time)
- coding becomes increasingly complex
- where is the manual?

when will lavaan support XXX?

- software package XXX can do XXX; why can't lavaan do this?
- lavaan is slow; when will you speed it up? I tried to fit this model with $P = 10,000$ variables, and it did not work
- when will you implement our new method in lavaan?
  - please read our (accepted/submitted) paper carefully
  - once it is done, add us as authors/contributors of the lavaan package
- person XXX volunteers to implement feature XXX:
  - (1 year later): sorry, something else came up, no time!
  - (1 year later): did I say I would do this?
  - oh, I decided to create my own package instead
  - sorry, I am no longer working in academia
- but I am guilty too: numerous times, I 'promised' that I would implement a feature, but never found time to do it

responsibility

- lavaan is increasingly used outside academia
- this results in mails that look like this:

> *I am a XXX working at XXX, a pharmaceutical company headquartered in XXX.*
>
> *We use 'lavaan' for some of our work, including analyses that are provided to health authorities (such as the FDA and EMA). These analyses take place in highly controlled environments — servers that run a validated version of R.*
>
> *We are always working to increase confidence in our software for hi-stakes decisions. To this extent, I am writing to inquire with you on the validation of the 'lavaan' package.*

- how do we know that lavaan works correctly?

six ways to report a potential bug

This is lavaan 0.6-21
lavaan is FREE software! Please report any bugs.

1. identify and document the bug, inform me, and send code that fixes the bug (or show exactly the line in the source code where the bug can be found)
2. identify and document the bug, and inform me (so I can investigate)
3. claim there might be a bug, but provide no reproducible example (reprex)
4. claim there is a bug, but provide no reproducible example (reprex)
5. write a paper, showing/claiming that 'lavaan is wrong' and just before submitting, send me the draft and ask for comments
6. write and publish a paper, showing/claiming that 'lavaan is wrong'; then send me an email saying: "we believe you may find this paper interesting"

why do we keep doing this?

- why stop:
  - no funding for software development (in Belgium)
  - no support at the faculty/university level: writing software is not in my job description; but being a department chair is ...
  - a few (lavaan) users are not very friendly
  - asking users/companies for a donation (since 2023) ... did not work
- why continue:
  - it is (for me) a way to learn about SEM, numerical techniques, statistics, mathematics, ...
  - it feels more useful than writing yet another paper
  - you meet interesting people
  - open-source (statistical) software is too important
- but at some point, others will have to take over

the future

- big news from December 2025: the Dutch NWO awarded an "Open Science Infrastructure" grant to the project:

  > "Infra-Structural Equation Modelling: Sustainably Scaling an Indispensable Open-Source Theory-Testing Infrastructure"

  led by a consortium of 11 Dutch professors (PI: Terrence Jorgensen)

- (I am involved as an external consultant)

- a big part of this grant is to write lavaan 2.0

- if all goes well, at the end of this project (4 years), the consortium will take over the day-to-day maintenance of lavaan

- long term: we are thinking of setting up a (not-for-profit) company that may provide lavaan/SEM-related 'services' (consultancy, teaching, customization, ...) to ensure sustainability

plans/ideas for lavaan 2.0

- lavaan becomes an interface to communicate with different engines, back-ends, packages, ...
  - input: model is described using lavaan syntax
  - computations: either internal or using external packages
  - output: results are shown in lavaan style
- core lavaan library in C++
  - R interface, but also Python/Julia/Matlab/...interface
- multiple engines, problem-specific
- new features:
  - short term: none
  - medium term: interfacing with other packages
  - long term: no idea yet

thanks!

- all the lavaan users
- teachers using lavaan
- researchers using lavaan to advance the field of SEM
- code contributors, package developers, the lavaan ecosystem
- lavaan discussion group volunteers
- workshop organizers
- seminar organizers
- ...
- please keep on using open-source software!