Thinking About and Approaching Statistics

Waylon Howard, PhD

Slides available at http://tinyurl.com/pharmerit-slides

PDF slides at http://tinyurl.com/pharmerit-pdf

Global Director of Biostatistics and Data Analytics

ConcertAI

@waylon-howard
@wwwaylon
wwwaylon.github.io

Education

Ph.D., Quantitative Psychology
University of Kansas (2012);
Advisor: Todd D. Little

Currently

Lead a 10-person team
Oncology focused RWE analytics

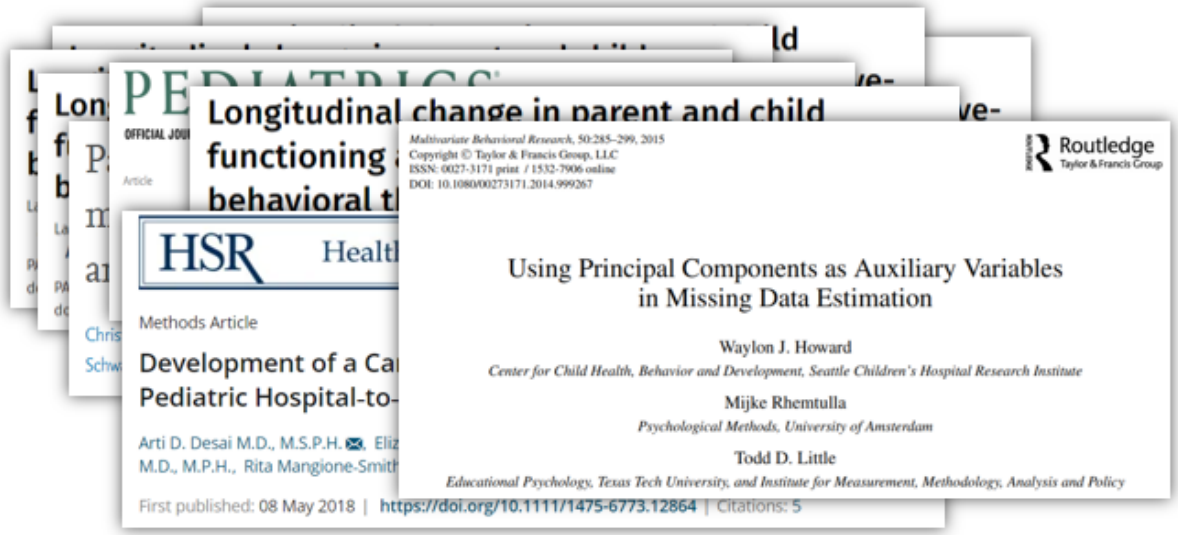
- Clinical: EMR curation
- Humanistic: PRO data
- Economic: claims

SymphonyAI acquisition

- More than 60% growth
- Developed 15+ SOP, WI, GD

Table of Contents

Interdisciplinary-oriented collaborator



My work has appeared in 22 different peer-reviewed journals garnering 518 citations, with an h-index of 13 and an i-10 index of 14.

Grants and contracts

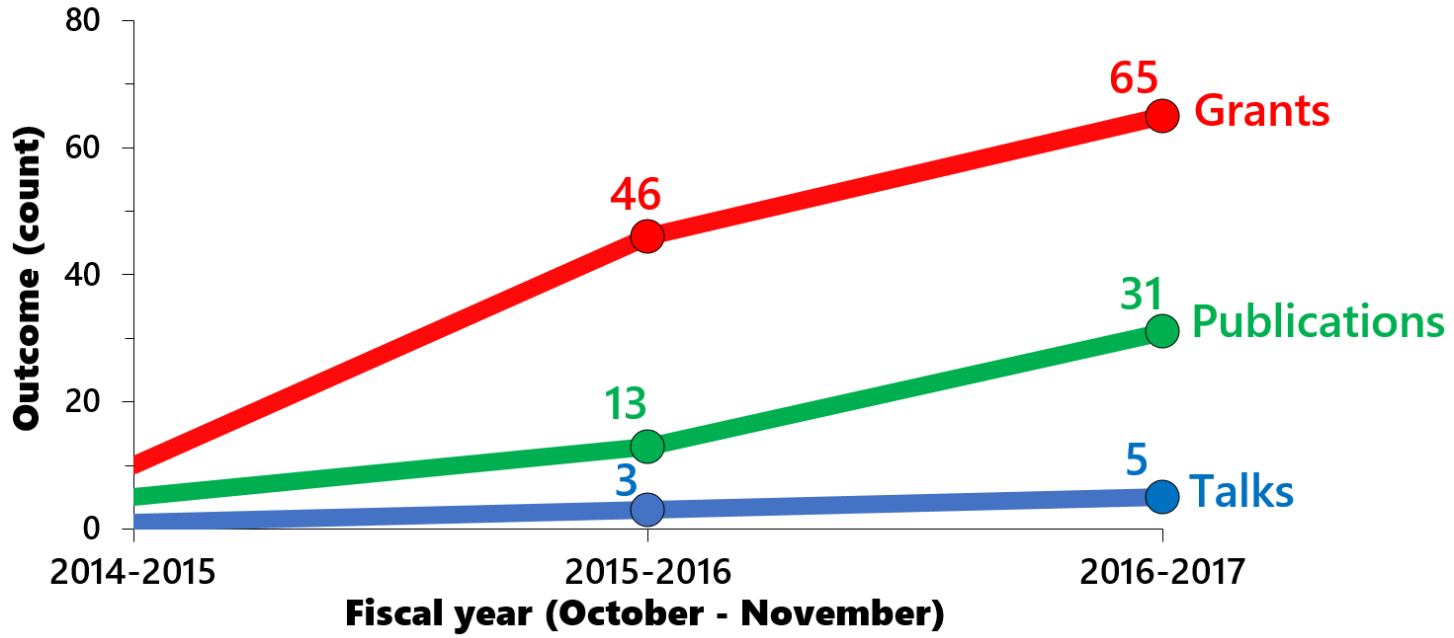
I have directly collaborated with researchers to attain **more than $5 million** in new research funding and developed considerable experience applying diverse best-practice methods to complex problems in new areas.

Training

Organized, managed, and taught an 8-course advanced methodology workshop for faculty and graduate students (N = 60) and a 6-course basic methodology training series for residents (N = 150).

Mentoring

The coalescence of a new vision for team leadership within the organization that resulted in new funding, partnerships, and additional value offered by the team.

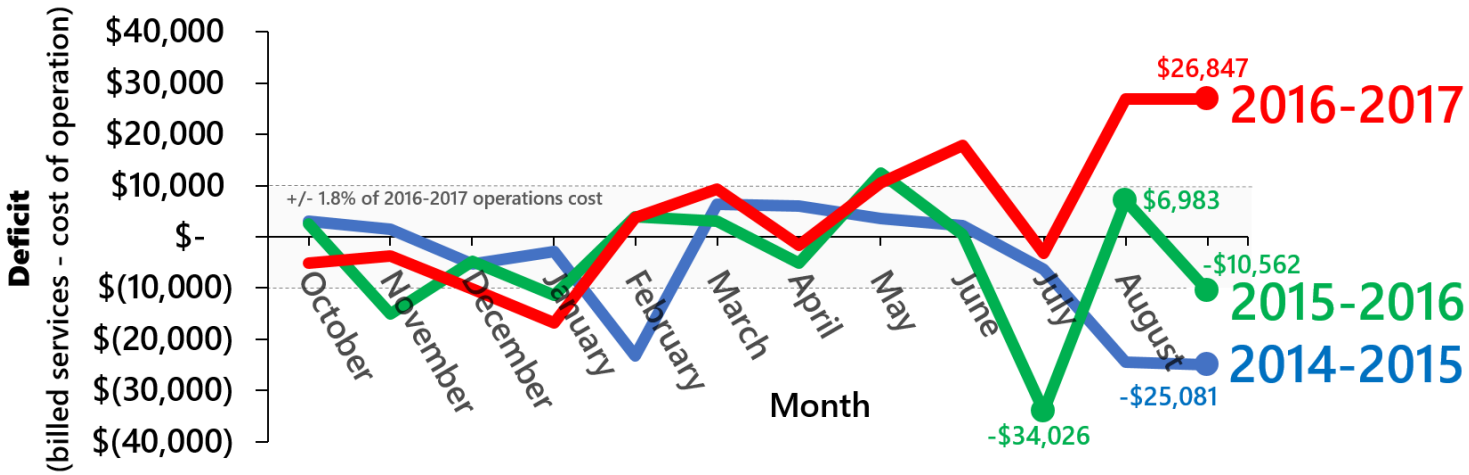


Is this change *significant*?

Different question:

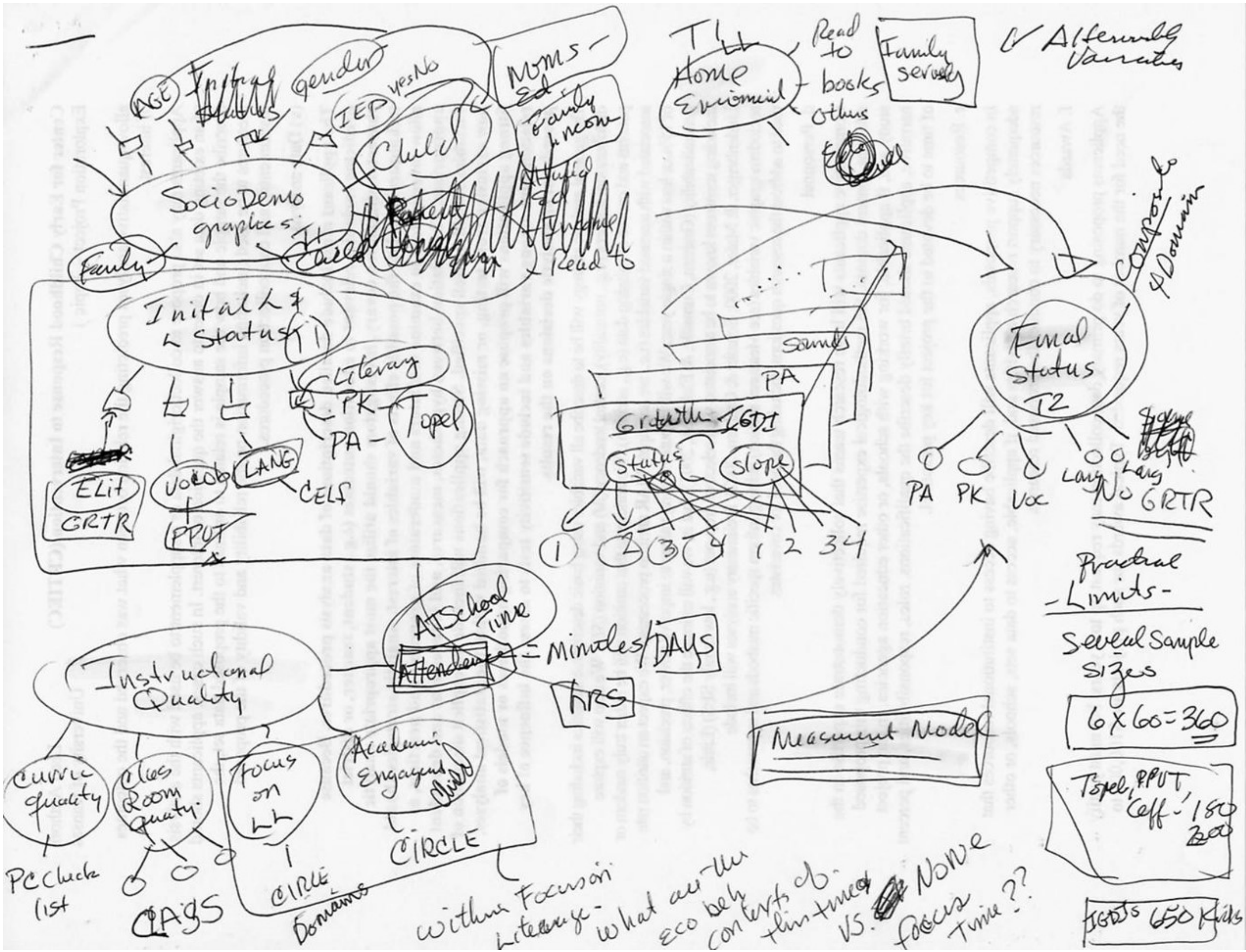
Is this change *sustainable*?

Remaining sustainable

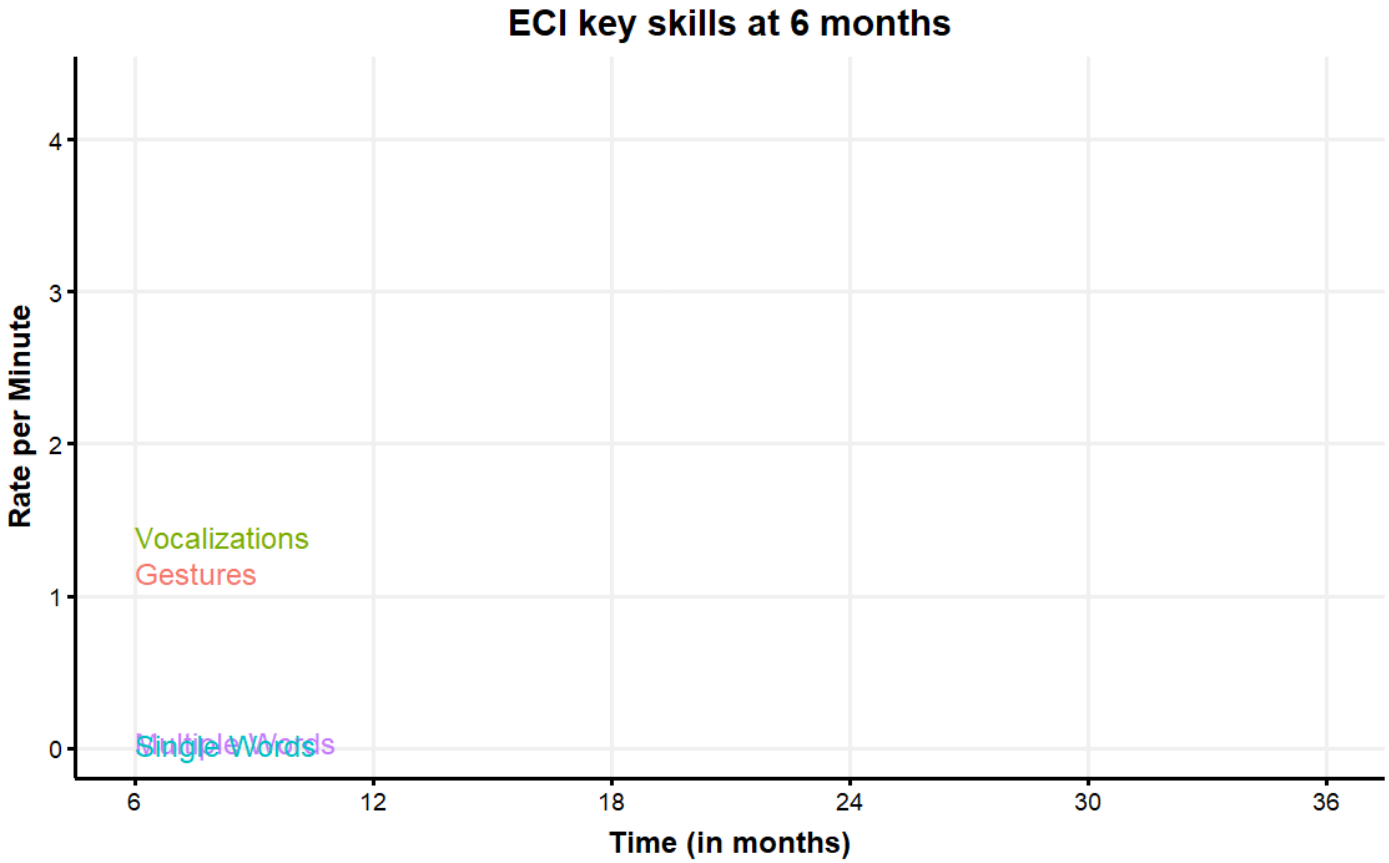


- Services billed increased by 38.6%
- Team averaged 78.3% billable time
- Compound annual growth rate +9.2%

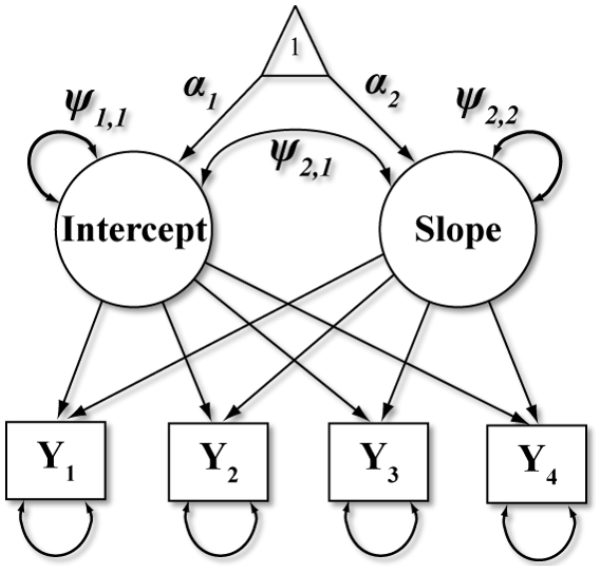
Motivating example


AGE
gender
IEP
yes No
Child
MOMS Ed
Family income
SocioDemographics
Family
Read to
Home Environment
Read to
books
others
Family services
Alternate Variables
Initial Status T1
Literacy
PK
PA
Topel
ELit
GRTR
Vocab
LANG
PPVT
CELF
Scores
PA
Growth IGDI
Status
Slope
1 2 3 4 1 2 3 4
Final Status T2
Composite 4 Domain
PA PK Voc Lang
Practical Limits
Several Sample sizes
6 X 60 = 360
Measurement Model
At School Time
Attendance
Minutes/DAYS
HRS
Instructional Quality
Curric quality
Class Room quality
Focus on LL
Academic Engaged Child
CIRCLE
CLASS
PC check list
Domains
within Focus on Literacy
what are the eco beh contexts of this time
vs. None Focus Time??
IGDIs 650 Kids

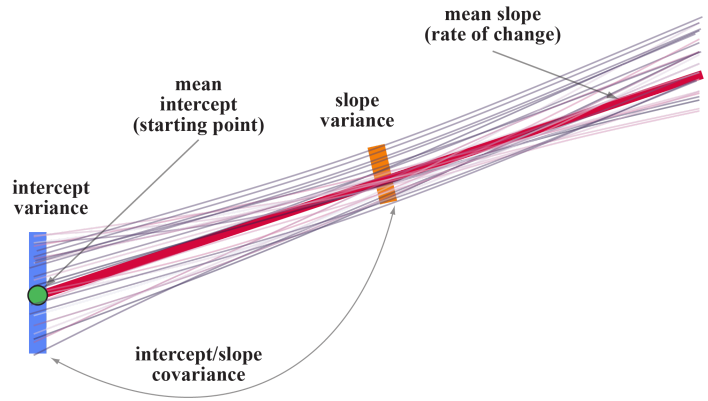
First motivating example for today



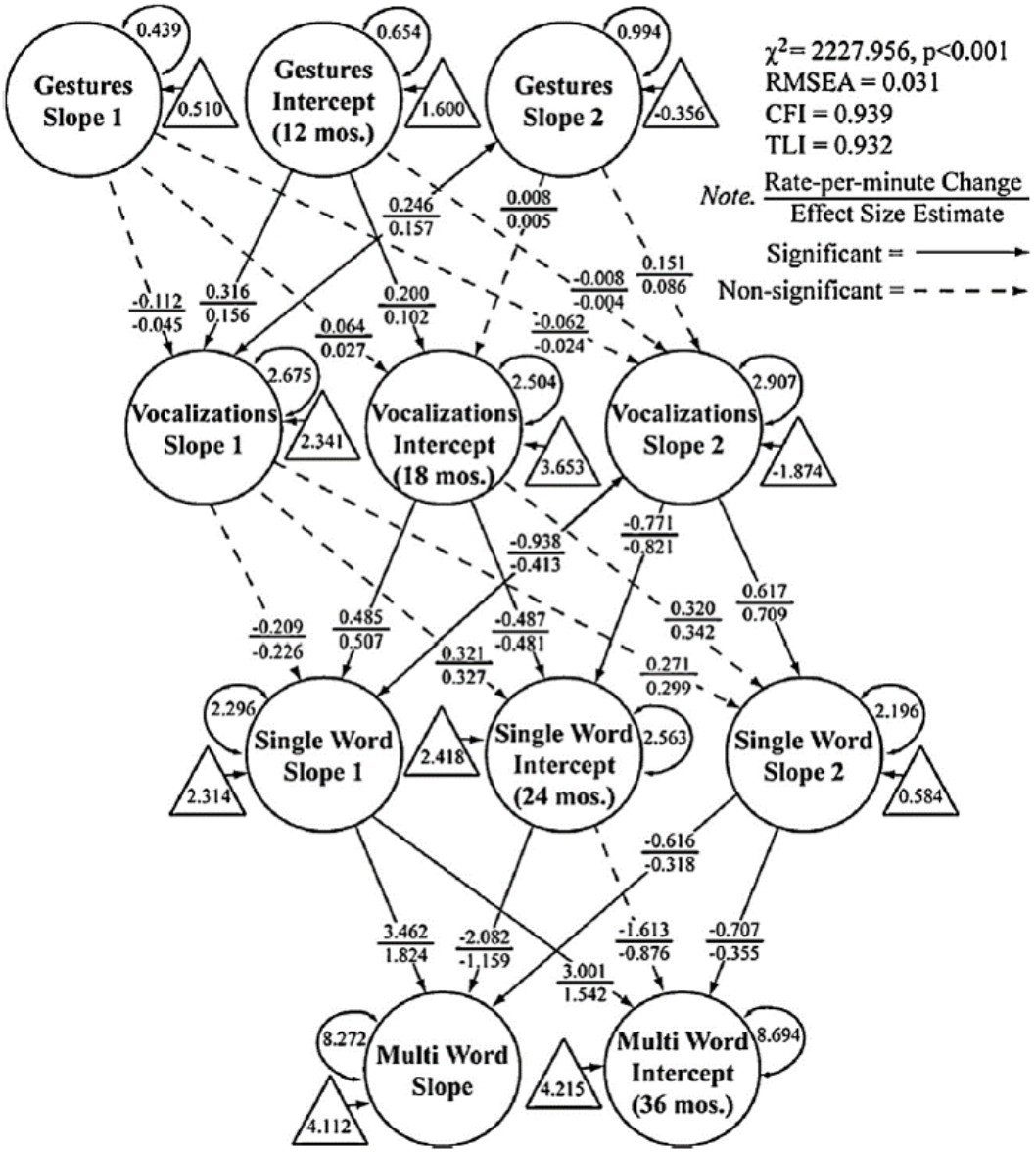
Exemplary LGM model



Path diagram



Parameter illustration

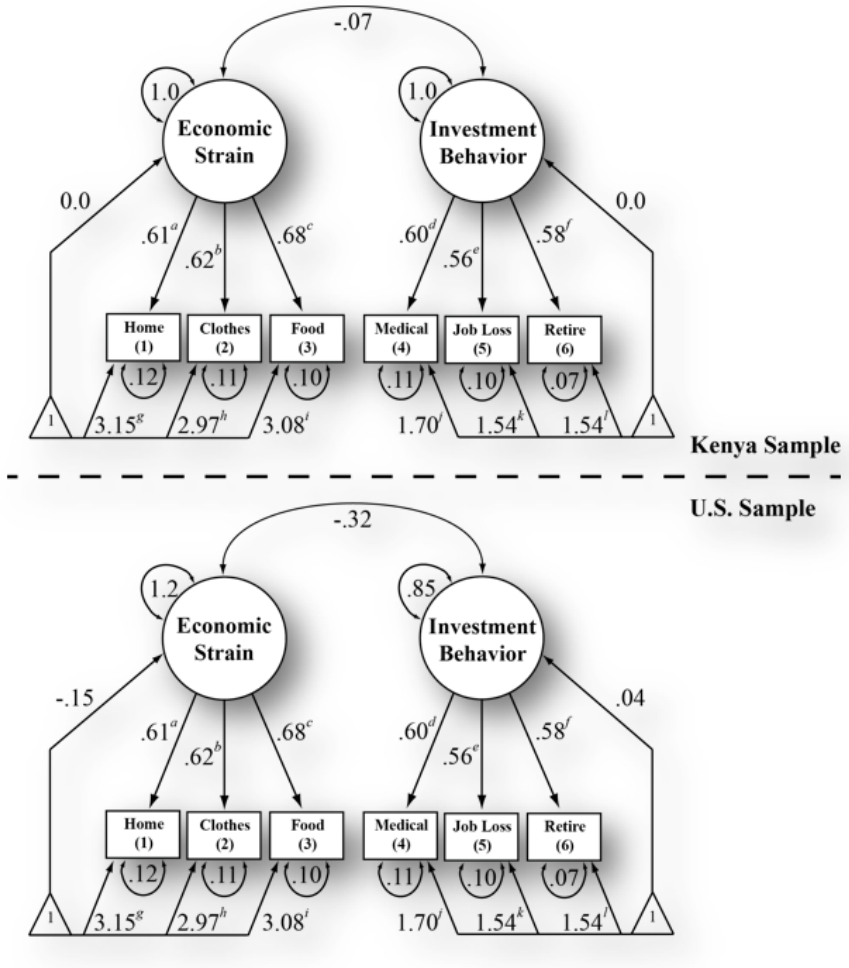

Gestures Slope 1
Gestures Intercept (12 mos.)
Gestures Slope 2
Vocalizations Slope 1
Vocalizations Intercept (18 mos.)
Vocalizations Slope 2
Single Word Slope 1
Single Word Intercept (24 mos.)
Single Word Slope 2
Multi Word Slope
Multi Word Intercept (36 mos.)
0.439
0.654
0.994
0.510
1.600
-0.356
2.675
2.504
2.907
2.341
3.653
-1.874
2.296
2.196
2.314
2.418
2.563
0.584
8.272
8.694
4.112
4.215
χ^2= 2227.956, p<0.001
RMSEA = 0.031
CFI = 0.939
TLI = 0.932
Note. Rate-per-minute Change / Effect Size Estimate
Significant =
Non-significant =
0.246 / 0.157
0.008 / 0.005
-0.112 / -0.045
0.316 / 0.156
0.200 / 0.102
-0.008 / -0.004
0.151 / 0.086
0.064 / 0.027
-0.062 / -0.024
-0.938 / -0.413
-0.771 / -0.821
-0.209 / -0.226
0.485 / 0.507
-0.487 / -0.481
0.320 / 0.342
0.617 / 0.709
0.321 / 0.327
0.271 / 0.299
-0.616 / -0.318
3.462 / 1.824
-2.082 / -1.159
3.001 / 1.542
-1.613 / -0.876
-0.707 / -0.355

Highlights: Measurement invariance

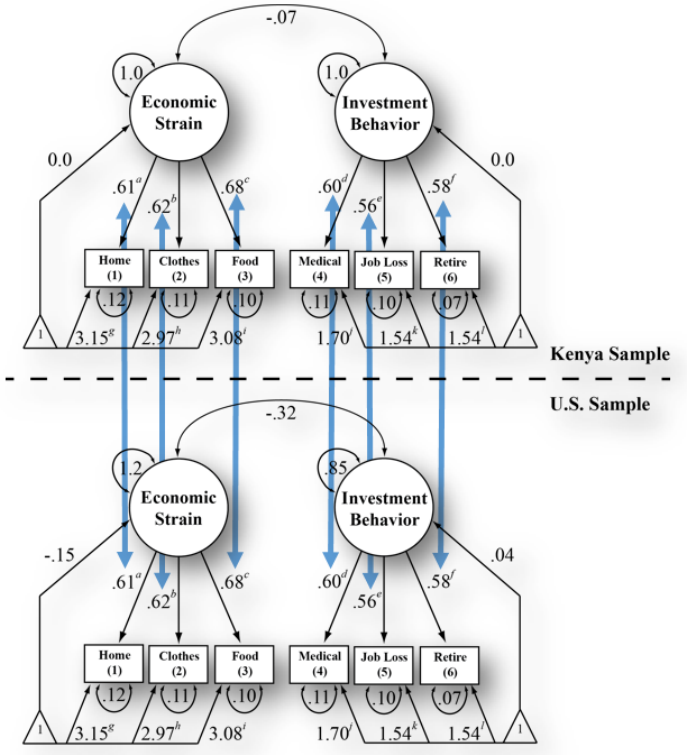
Why?

To ensure we are measuring the same constructs across groups and time - allows for structural invariance testing

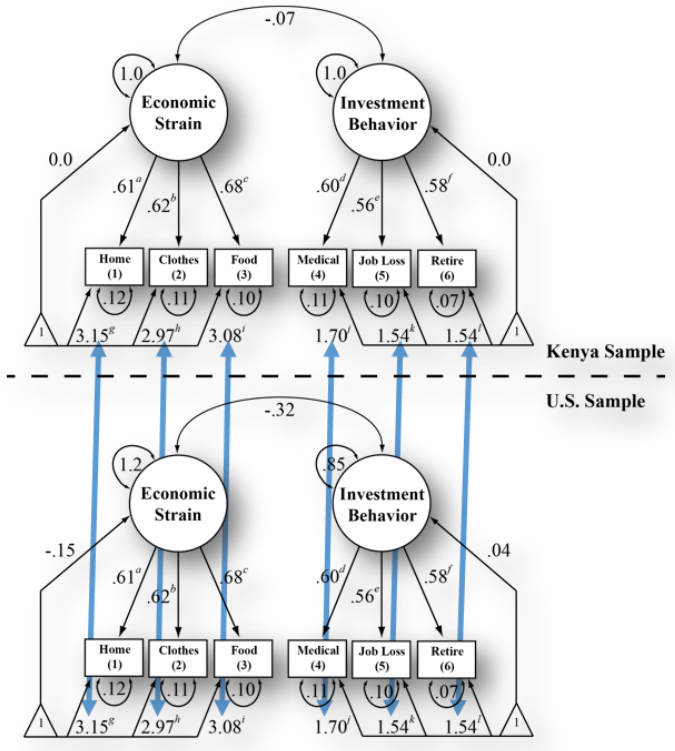
1. Configural invariance



2. Weak invariance

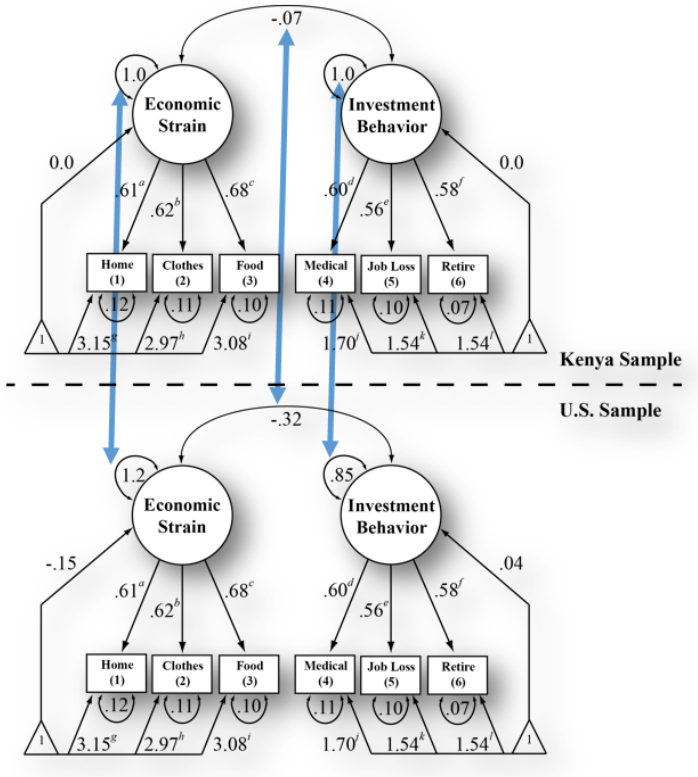


3. Strong invariance

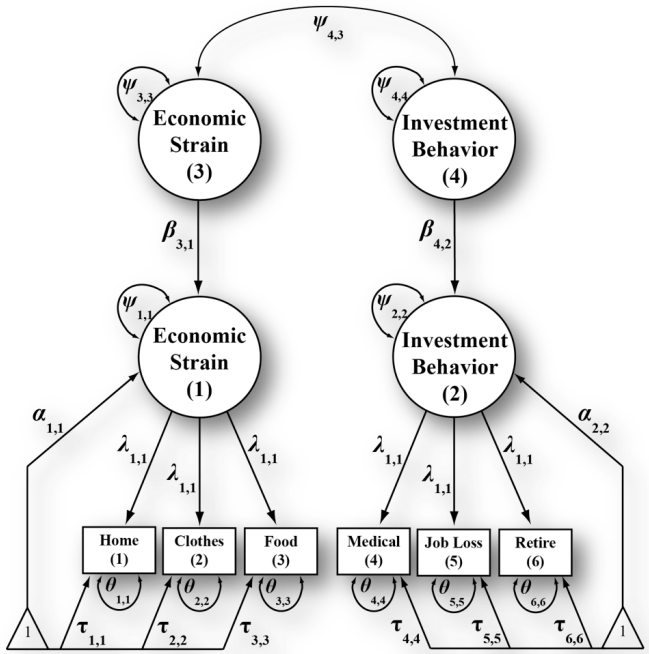


Structural invariance

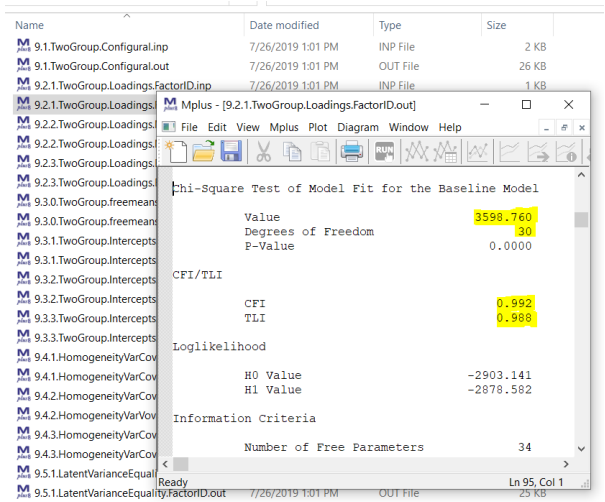
Latent correlations



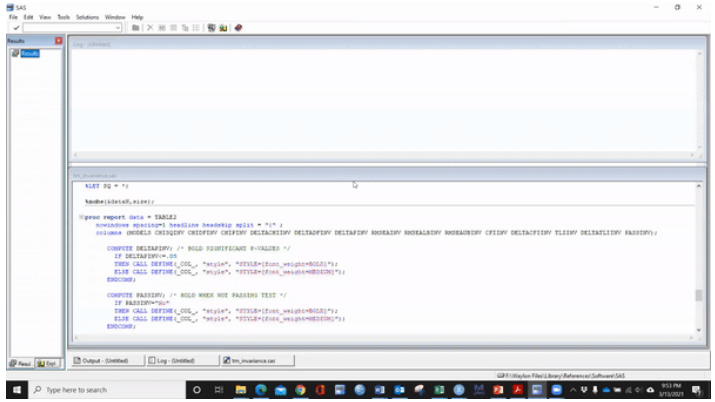
Phantom variables



A note about efficiency tools

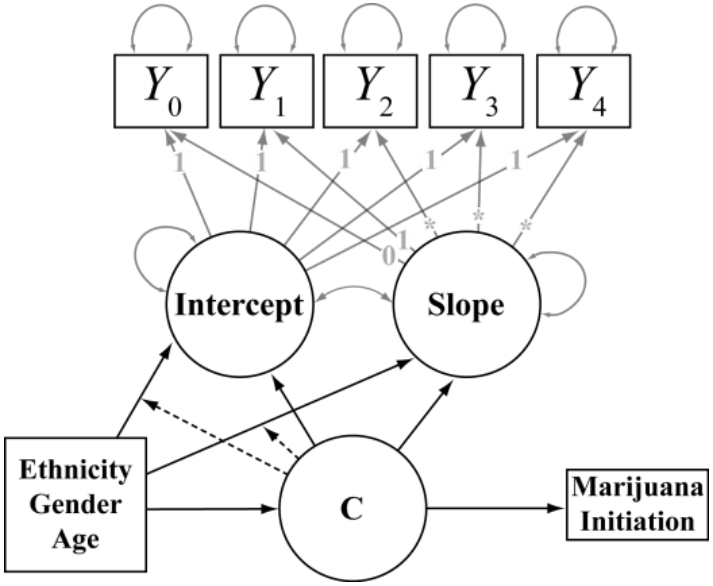


Numerous Mplus output files



Example model fit table automation

A note about organizational tools



Σ_k	# of classes (K)	LL	$npar$	AIC	BIC	CAIC	AWE	LRTS	Adj LMR p-value	$BF_{K,K-1}$	cmP_K
	1	-1079.65	4	2167.29	2179.22	2183.22	2211.16	11.96	---	4.46	---
Most Restricted Model	2	-1073.66	7	2161.33	2182.21	2189.21	2238.10	13.39	0.014	2.18	0.1690
Class-invariant, diagonal	3	-1066.97	10	2153.94	2183.77	2193.77	2263.61	2.80	0.027	435.03	0.0775
$\Sigma_k = \Sigma$ Variances freely	4	-1065.57	13	2157.14	2195.92	2208.92	2299.71	5.09	0.199	138.16	0.0002
estiamted BUT constrained	5	-1063.02	16	2158.04	2205.78	2221.78	2333.52	---	0.709	---	0.0000
across groups - covariances				---	---	---	---	---	---	---	---
fixed to zero (**psidiagciv**.inp)				---	---	---	---	---	---	---	---
	1	-1079.65	4	2167.29	2179.22	2183.22	2211.16	23.65	---	1.88	---
Class-varying, diagonal Σ_k	2	-1067.82	9	2153.64	2180.49	2189.49	2252.34	10.29	0.002	1499.69	0.4000
Variances freely estimated BUT	3	-1062.67	14	2153.35	2195.12	2209.12	2306.89	5.35	0.160	1471.76	0.0003
covariances fixed to zero	4	-1060.00	18	2156.00	2209.70	2227.70	2353.41	---	0.000	---	0.0000
(**psidiagcvar.inp**)				---	---	---	---	---	---	---	---
				---	---	---	---	---	---	---	---
				---	---	---	---	---	---	---	---
	1	-1077.56	5	2165.11	2180.03	2185.03	2219.95	9.07	---	18.94	---
Class-invariant,	2	-1073.02	8	2162.05	2185.91	2193.91	2249.78	15.08	0.312	0.94	0.0266
unrestricted $\Sigma_k = \Sigma$ Variances	3	-1065.49	11	2152.97	2185.79	2196.79	2273.61	3.62	0.026	289.00	0.0283
freely estimated BUT	4	-1063.68	14	2155.35	2197.12	2211.12	2308.89	4.61	0.603	176.34	0.0001
covariances equated across	5	-1061.37	17	2156.75	2207.47	2224.47	2343.19	---	0.221	---	0.0000
groups (**psinondiagcinv.inp**)				---	---	---	---	---	---	---	---
				---	---	---	---	---	---	---	---

a transparent, traceable tool

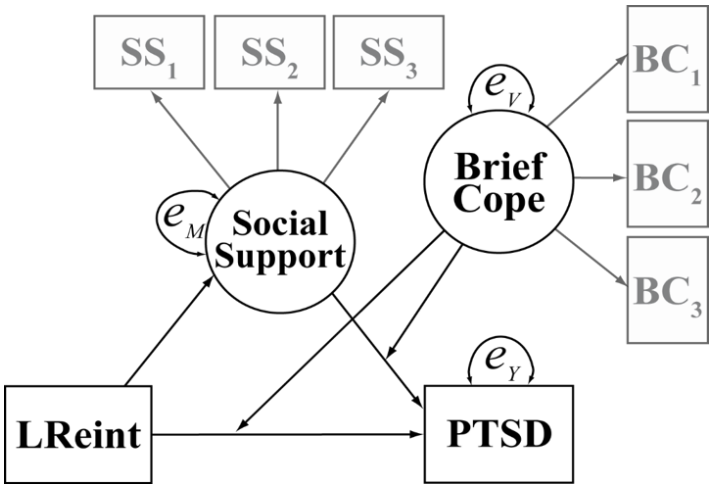
a complex, iterative process...

Highlights: moderated mediation

Why?

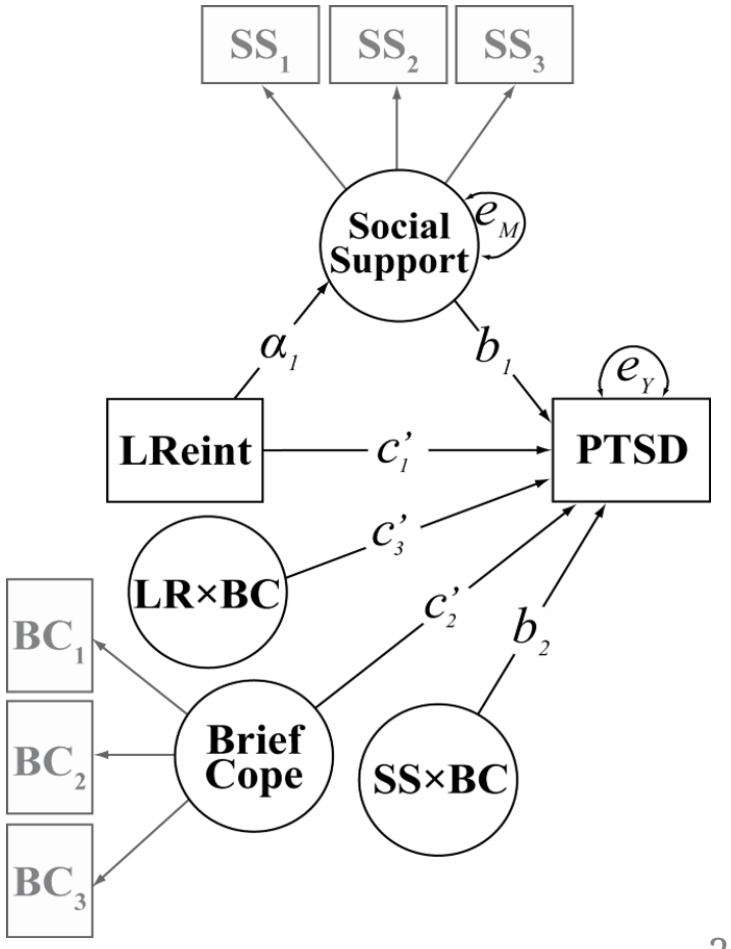
We often want to know how (mediation) and when (moderation) predictive relations hold or are strong versus weak

Conceptual diagram



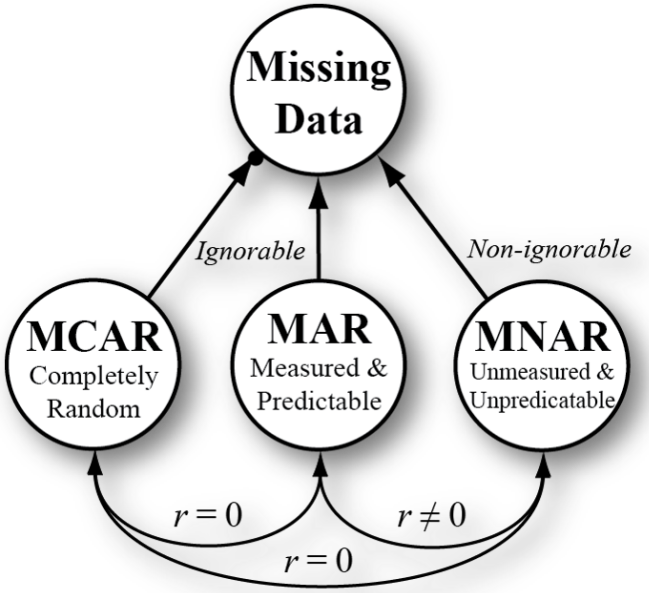
See Full-text available

Statistical diagram



Highlights: Missing data and power

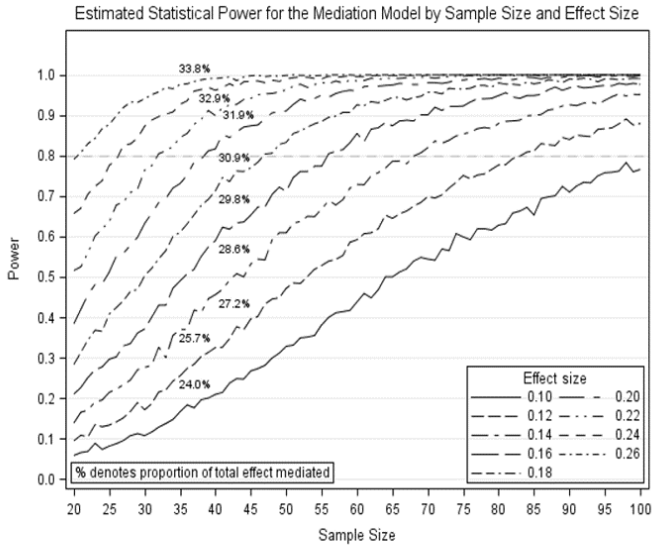
Missing Data Analysis



Simulation in MBR Full-text

Reporting practices Full-text

Monte Carlo power simulations



Software wrangling

Statistical library and templates

Summary

1. Leading technical experts for the past 10 years

2. Extensive real-world experience applying diverse best-practice methods to complex problems in new areas

3. Readily pivot from the strategic to hands on

4. Thrive in fast-paced, changing work environment

Thanks!

Any further questions?

- Slides created via the R package xaringan by Yihui Xie
- Slides' source code at https://wwwaylon.github.io/assets/ph-2021/
- R code from throughout the slides as an R script as slide_code.R

Appendix

Previous research appointments

Biostatistics Manager
Dir. of Biostatistics
Dir. of Biostatistics Core
Faculty Res. Methodologist
Dir. of Res. & Evaluation
Senior Res. Data Analyst
KU Quantitative Analyst

- Position details here

Full-text research content

- Peer-reviewed research articles
- Talks, workshops, posters

Exemplar research tools

- Consultant-based effort estimator
- Timeline visualizations for project management and communication
- Rmarkdown automation of CV, NIH biosketch, NSF biosketch, etc.

Example standard process development

Training	Document ID	Document Title
Course-015	GD-OSS015-R01	Statistical Programming Best Practices
Course-016	GD-OSS016-R01	Analysis Datasets Results Verification
Course-017	GD-OSS017-R01	Reusable Code Validation
Course-020	GD-OSS020-R01	Procedures for Annotation and QC of TLFs
Course-001	CHKLST-OSS001	Source Code Version Control
Course-002	CHKLST-OSS002	Programming Quality Control Checklist
Course-003	CHKLST-OSS003	Disk Space Management Checklist
Course-004	CHKLST-OSS004	AWS Import/export Checklist
Course-005	CHKLST-OSS005	Data Delivery Checklist
Course-006	CHKLST-OSS006	Chemo Master List
Course-006	WI-OSS001	De-identification and Data Transfer

Software summary (not all-inclusive)

- SAS 9.4 (STAT, IML)
- Mplus 8.4 (base, mixture, multilevel)
- IBM SPSS statistics 25 (base, missing values, AMOS)
- R (data manipulation: dplyr, tidyr, stringr, RMySQL, RSQLite; data visualization: ggplot2, htmlwidgets; reporting: shiny, rmarkdown; analysis: lavaan, psych, lme4/nlme, survival)

Methodology interest summary

- Design and measurement issues in longitudinal research, panel designs, latent growth curve analysis, latent class and finite mixture modeling, multi-level SEM with longitudinal data, moderation and mediation, missing data analysis and power, and measurement invariance testing.