Modeling approaches for cross-sectional integrative data analysis

Evaluations and recommendations

Kenneth Tyler Wilcox & Lijuan Wang

Department of Psychology, University of Notre Dame

IMPS, 21 July 2021

What is Integrative Data Analysis?

Advantages

Current Practice

Participant-Level and Study-Level Effects

Integrative Data Analysis (IDA)

> Integrative data analysis (IDA) simultaneously analyzes the *participant-level* data from multiple studies (Curran & Hussong, 2009)

- Also known as
  - individual participant meta-analysis (Cooper & Patall, 2009)
  - individual patient data meta-analysis (Stewart & Tierney, 2002)
  - mega-analysis (McArdle et al., 2009)
  - data fusion (Marcoulides & Grimm, 2017)

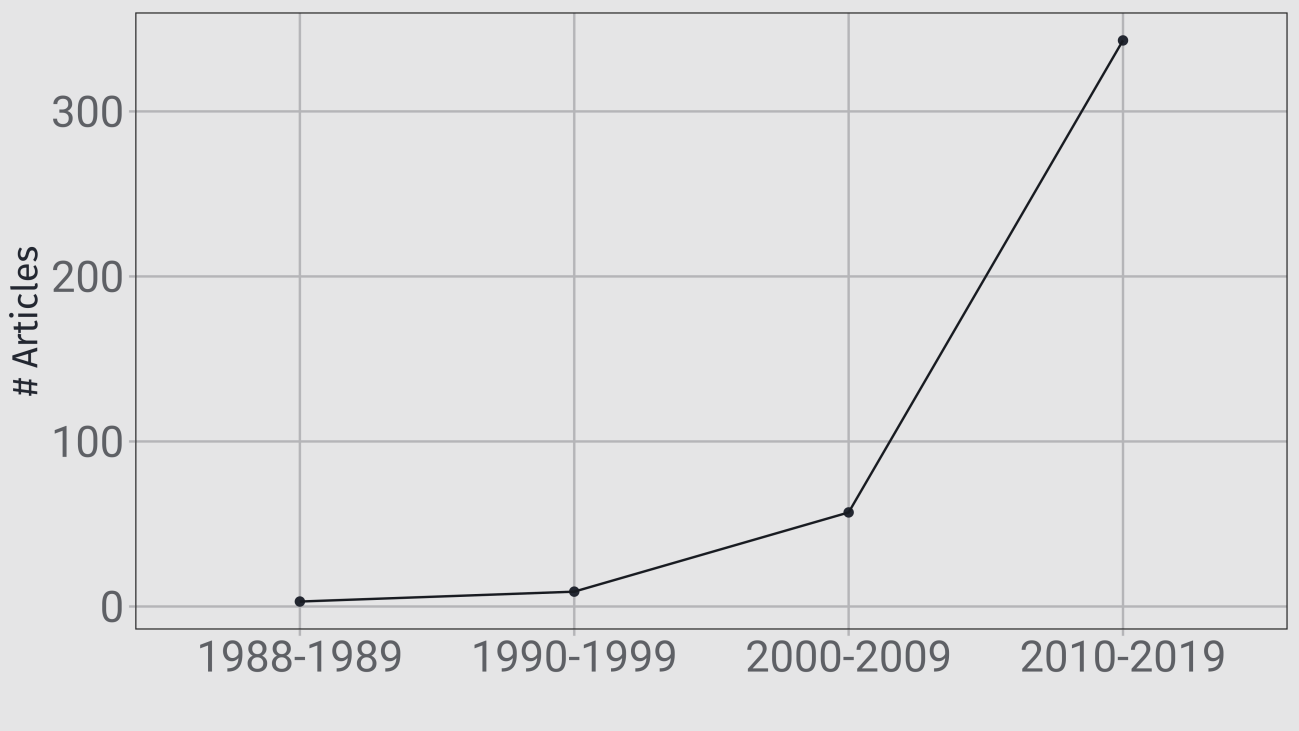
Advantages of IDA

- Use of multiple samples introduces and allows modeling of between-sample heterogeneity
- Directly assess the replicability of effects across studies and populations
- Can fit more complex models and answer new research questions
- Longitudinal analysis of longer timespans is often possible
- Improved harmonic measurements

(Bauer & Hussong, 2009; Curran et al., 2018; Curran & Hussong, 2009; Marcoulides & Grimm, 2017; McArdle et al., 2009; Stewart & Tierney, 2002)

Current Practice of IDA in Psychology

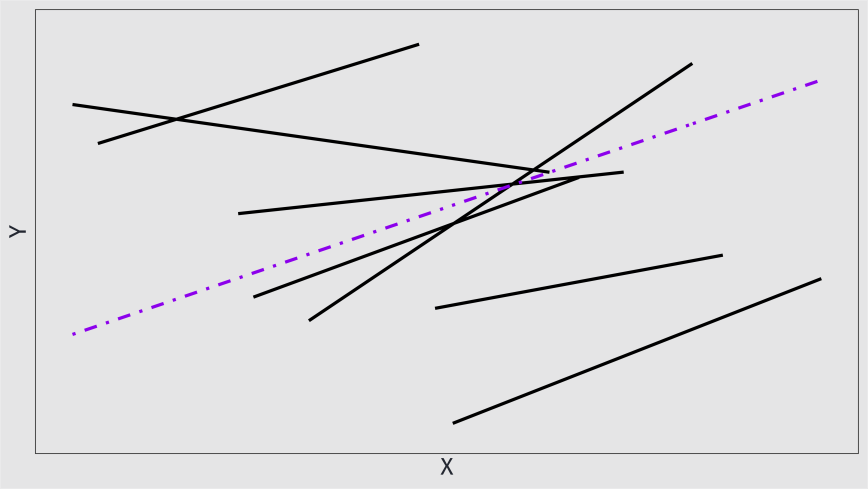
- PsycINFO literature search: 1988--2020



- 91% of 421 articles used fixed-effects models; minimal disaggregation

Participant-Level and Study-Level Effects

Participant-Level Effects per Study



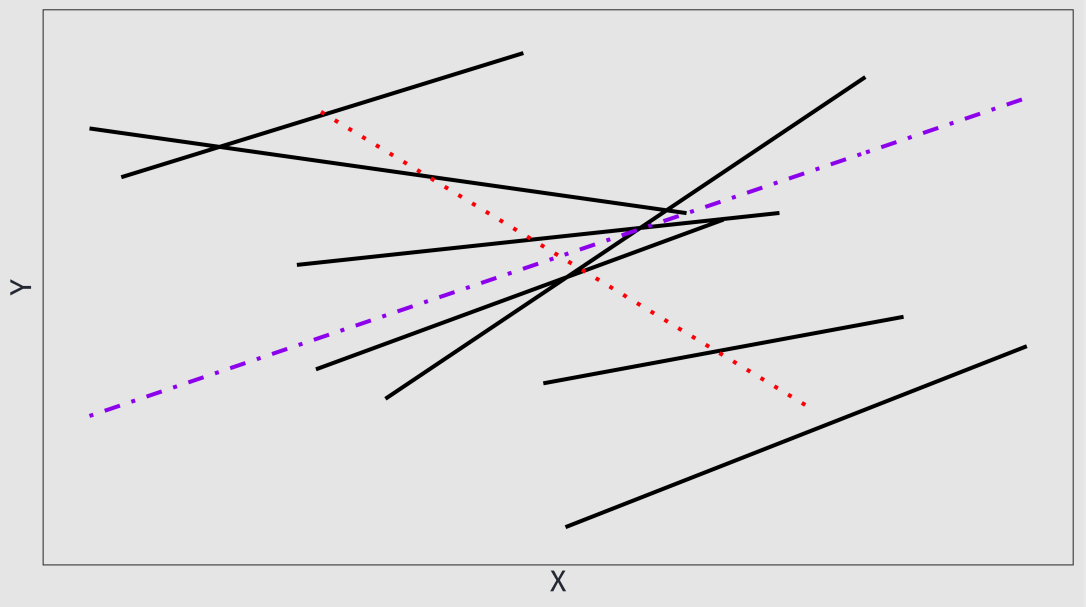
Average participant-level effect of X on Y: γ_W (dot-dashed/purple line)

Variability of intercepts $\rightarrow \sigma^2_{u_0}$

Variability of slopes $\rightarrow \sigma^2_{u_1}$

Participant-Level and Study-Level Effects

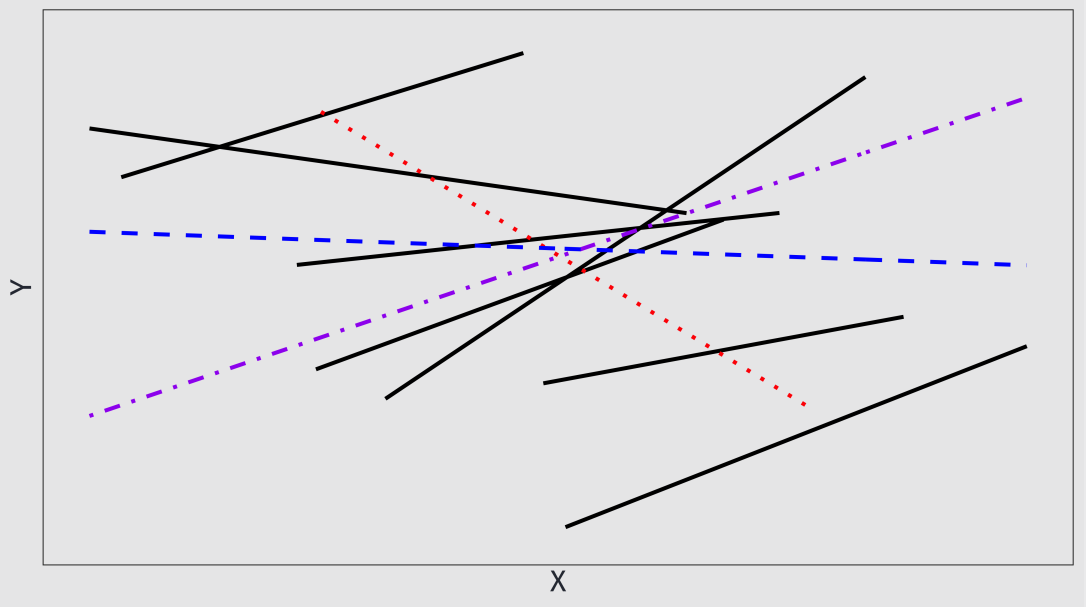
Study-Level Effect



Study-level effect of $\bar{X}$ on $\bar{Y}$: γ_B (dotted/red line)

Participant-Level and Study-Level Effects

Failure to Disaggregate



Aggregated effect: γ_A (dashed/blue line)

"An uninterpretable blend" (Raudenbush & Bryk, 2002, p. 138) of γ_W and γ_B

Research Questions

1. What Models Can Disaggregate Participant- and Study-Level Effects?

2. How Do We Account for Between-Study Heterogeneity?

3. What Methods Work in IDA Small Sample Scenarios?

IDA Models

Aggregated Regression

Disaggregated Regression

Study-Specific Coefficients Regression

Fixed-Slope Multilevel Model

Random-Slopes Multilevel Model

Aggregated vs. Disaggregated Regression

Aggregated Regression

$$y_{ij} = \gamma_{00} + \gamma_A x_{ij} + e_{ij}$$
$$e_{ij} \sim \mathrm{N}\left(0, \sigma_e^2\right)$$

- γ_A conflates participant- and study-level fixed effects as a weighted function of the intrastudy correlation
  - $\gamma_A = (1 - \lambda)\gamma_W + \lambda\gamma_B$
- Popular in application

Disaggregated Regression

$$y_{ij} = \gamma_{00}^* + \gamma_B \bar{x}_j + \gamma_W \left(x_{ij} - \bar{x}_j\right) + e_{ij}$$
$$e_{ij} \sim \mathrm{N}\left(0, \sigma_e^2\right)$$

- γ_B: study-level fixed effect
- γ_W: average participant-level fixed effect

Sources of between-study heterogeneity are *ignored*

(Hamaker & Muthén, 2020; Neuhaus & Kalbfleisch, 1998; Raudenbush & Bryk, 2002)

Study-Specific Coefficients Regression

- Extends disaggregated regression model to model **mean heterogeneity**

$$y_{ij} = \sum_{k=1}^{J} \gamma_{0k} I(k = j) + \gamma_W x_{ij} + e_{ij}$$
$$e_{ij} \sim \mathrm{N}\left(0, \sigma_e^2\right)$$

- *Cannot* include study-level effect of $\bar{x}_j$
- Accounts for between-study heterogeneity in study outcome means $\bar{y}_j$

(Curran & Hussong, 2009)

Fixed-Slope Multilevel Model

- Extends the disaggregated regression model to model **mean heterogeneity**

$$
\begin{aligned}
\textbf{Level 1: } y_{ij} &= \beta_{0j} + \beta_{1j}\left(x_{ij} - \bar{x}_j\right) + e_{ij} \\
\textbf{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_B \bar{x}_j + u_{0j} \\
\beta_{1j} &= \gamma_W \\
\begin{bmatrix} \vec{e}_j \\ u_{0j} \end{bmatrix} &\sim \mathrm{N}\left(\begin{bmatrix} \vec{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 \vec{I}_{n_j} & 0 \\ 0 & \sigma_{u_0}^2 \end{bmatrix}\right)
\end{aligned}
$$

- $\sigma_{u_0}^2$: between-study variance in study conditional means
- More parsimonious than the SSC regression model at the cost of a distributional assumption

Random-Slopes Multilevel Model

- Extend the fixed-slope MLM to incorporate **heterogeneity in (1) means and (2) participant-level effects**

$$
\begin{aligned}
\textbf{Level 1: } y_{ij} &= \beta_{0j} + \beta_{1j}\left(x_{ij} - \bar{x}_j\right) + e_{ij} \\
\textbf{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_B \bar{x}_j + u_{0j} \\
\beta_{1j} &= \gamma_W + u_{1j} \\
\begin{bmatrix} \vec{e}_j \\ u_{0j} \\ u_{1j} \end{bmatrix} &\sim \mathrm{N}\left(\begin{bmatrix} \vec{0} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 \vec{I}_{n_j} & 0 & 0 \\ 0 & \sigma_{u_0}^2 & \sigma_{u_{01}} \\ 0 & \sigma_{u_{01}} & \sigma_{u_1}^2 \end{bmatrix} \right)
\end{aligned}
$$

- $\sigma_{u_1}^2$: between-study variance in participant-level effects

RQ1 and RQ2: Disaggregation and Heterogeneity

Table 1

Overview of Five Models for IDA

Model	Equations	γ_B	γ_W	γ_A	$\sigma^2_{u_0}$	$\sigma^2_{u_1}$
A LR	$y_{ij} = \gamma^*_{00} + \gamma_A x_{ij} + e_{ij}$ $e_{ij} \sim \mathrm{N}(0, \sigma^2_e)$			✓	= 0	= 0
D LR	$y_{ij} = \gamma_{00} + \gamma_W (x_{ij} - \overline{x}_j) + \gamma_B \overline{x}_j + e_{ij}$ $e_{ij} \sim \mathrm{N}(0, \sigma^2_e)$	✓	✓		= 0	= 0
SSC LR	$y_{ij} = \sum_{k=1}^{J} \gamma_{0k} I(k = j) + \gamma_W x_{ij} + e_{ij}$ $e_{ij} \sim \mathrm{N}(0, \sigma^2_e)$		✓			= 0
FS MLM	$y_{ij} = \beta_{0j} + \gamma_W (x_{ij} - \overline{x}_j) + e_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_B \overline{x}_j + u_{0j}$ $u_{0j} \sim \mathrm{N}\left(0, \sigma^2_{u_0}\right),\ e_{ij} \sim \mathrm{N}(0, \sigma^2_e)$	✓	✓		✓	= 0
RS MLM	$y_{ij} = \beta_{0j} + \beta_{1j} (x_{ij} - \overline{x}_j) + e_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_B \overline{x}_j + u_{0j}$ $\beta_{1j} = \gamma_W + u_{1j}$ $\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_{01}} & \sigma^2_{u_1} \end{bmatrix} \right),\ e_{ij} \sim \mathrm{N}(0, \sigma^2_e)$	✓	✓		✓	✓

RQ3: Small Sample IDA Methods and Performance

Underevaluated Impact of Variance Effect Sizes

Underevaluated MLM Degrees of Freedom Methods for IDA

Simulation Study Design

- Generated data from fixed-slope and random-slopes MLMs with 1,000 replications
- Unbalanced study sample sizes based on Hornburg et al. (2018)
- Set parameters using proportion of variance effect sizes (Rights & Sterba, 2019)
- Factors
  - Number of studies: 2, 3, . . ., 35
  - Average study sample size: 25, 51, 101
  - Effect size of γ_B: 0, "small", "medium"
  - Effect size of γ_W: 0, "small", "medium"
  - Effect size of $\sigma^2_{u_0}$: 0, "small", "medium"
  - Effect size of $\sigma^2_{u_1}$: 0, "small", "medium"
- Evaluated degrees of freedom (DF) methods in SAS Proc MIXED: Residual, Containment, Between-Within, Satterthwaite, Kenward-Roger

Testing γ_B Depends on DF Method and $\sigma^2_{u_0}$

- Type I error rate for study-level effect affected by degree of mean heterogeneity
  - $\sigma^2_{u_0}$ needs to be modeled if $\sigma^2_{u_0} > 0$: FS MLM or RS MLM
  - Type I error rate depends on effect size of $\sigma^2_{u_0}$ and DF method
- Between-Within, Satterthwaite and Kenward-Roger DF worked well with at least 5-14 studies
  - For small $\sigma^2_{u_0}$, Satterthwaite DF needed fewer (6) studies
  - For medium $\sigma^2_{u_0}$, Kenward-Roger DF needed fewer (5) studies

Testing γ_W Depends on DF Method and $\sigma^2_{u_1}$

- Type I error rate for average participant-level fixed effect affected by degree of participant-level effect heterogeneity
  - $\boldsymbol{\sigma^2_{u_1}}$ needs to be modeled if $\boldsymbol{\sigma^2_{u_1} > 0}$: RS MLM
  - Type I error rate depends on effect size of $\boldsymbol{\sigma^2_{u_1}}$ and DF method
- Containment, Satterthwaite, and Kenward-Roger DF methods worked well with at least 4-15 studies
  - Previous research recommended Kenward-Roger DF
  - Containment DF needed fewer (5 or 6) studies (see also Ferron et al., 2009)

(Huang, 2016; Kenward & Roger, 1997; McNeish, 2017; McNeish & Stapleton, 2016; Morris et al., 2018)

Recommendations

Recommendations

- Disaggregate participant-level and study-level fixed effects
- Carefully consider and model sources of between-study heterogeneity
    - Failing to do so can yield incorrect type I error rates for one or both levels of fixed effects
- With a small number of studies, random-slopes MLM can yield accurate estimates and well-controlled type I error rates for both types of fixed effects
    - Appropriate degrees of freedom methods are critical
        - Kenward-Roger (1997) DF for study-level fixed effect
        - Containment DF for participant-level fixed effect
- Overall, MLM can be a viable option for IDA with even as few as six studies

Thanks!

kwilcox3@nd.edu

www.ktylerwilcox.me

Slides:

https://www.ktylerwilcox.me/talk/2021-imps-ida/

Paper:

Wilcox, K. T., & Wang, L. (In press). Modeling approaches for cross-sectional integrative data analysis: Evaluations and recommendations. *Psychological Methods*. https://doi.org/10.1037/met0000397