Integrated Data Analysis in Prevention Science

Presented by:
Rashelle J. Musci, Ph.D.
Johns Hopkins Bloomberg School of Public Health
INTEGRATED DATA ANALYSIS IN PREVENTION SCIENCE

Rashelle J. Musci, Ph.D.
Department of Mental Health
Land Acknowledgement

- This presentation was developed on stolen Nanticoke and Piscataway land.
- [https://native-land.ca/](https://native-land.ca/)

Redlining Acknowledgment

- This presentation was developed in a community impacted by redlining and gentrification.
- [https://dsl.richmond.edu/panorama/redlining](https://dsl.richmond.edu/panorama/redlining)
GOALS IN PREVENTION SCIENCE

- Prevention social, physical, mental health, and academic problems.
- Promote health and well-being
- Research, support, and implement evidenced based programming
Large scale RCTs are expensive and resource heavy, typically designed and powered for main outcome analyses.

Increasingly, the literature is pointing toward an emphasis on mechanisms and subgroups.

Researchers, practitioners, and policy makers are interested in rarer outcomes and exposures.
WHAT IS INTEGRATIVE DATA ANALYSIS?

- A set of techniques used to fit models to pooled or harmonized data
  - Allows us to capitalize on between-study heterogeneity
  - Increases power for mediation and moderation analyses
  - Allows for longitudinal models to be used across developmental periods
- To date, most IDA studies within mental health research have focused either on cohort studies or data from clinical trials with psychiatric medication.
Ultimate goal:

- Integrative data analysis is an important methodological tool in your prevention research toolbox.

Our motivating example:

- Early childhood prevention programming and suicidal behaviors
BACKGROUND

- Suicide is a significant public health concern, particularly among black youth (13 and below) who are more likely to die from suicide compared to their white peers.

- Early prevention programming has shown significant impacts on suicidal behaviors across the life course.

- We do expect that universal programming to have some effect heterogeneity.
  - Exploring this can be challenging within study, particularly with rare outcomes.
INTRODUCING THE DATA

Through the use of integrative data analysis (IDA) we harmonize extant school-based cluster randomized prevention trials with longitudinal follow up

- JHU Center for prevention & early intervention, 1st generation trial
- JHU Center for prevention & early intervention, 2nd generation trial
- JHU Center for prevention & early intervention, Paths to PAX trial
- Schools and Families Educating (SAFE) Children Study
- Fast Track Project
- Linking the Interests of Families and Teachers (LIFT) study
## DATASET INFORMATION

### Table 1. Basic Information about included Datasets

<table>
<thead>
<tr>
<th>Prevention Type</th>
<th>PIRC G1</th>
<th>PIRC G2</th>
<th>P2P</th>
<th>Fast Track</th>
<th>LIFT</th>
<th>SAFE Children</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Universal</td>
<td>Universal</td>
<td>Universal</td>
<td>Universal + Indicated</td>
<td>Universal</td>
<td>Universal</td>
</tr>
<tr>
<td><strong>Randomization Level</strong></td>
<td>School &amp; Classroom</td>
<td>Classroom</td>
<td>School</td>
<td>School</td>
<td>School</td>
<td>Classroom</td>
</tr>
<tr>
<td>Sample size</td>
<td>2,311</td>
<td>678</td>
<td>5,611</td>
<td>1,199</td>
<td>671</td>
<td>424</td>
</tr>
<tr>
<td>Male (%)</td>
<td>50%</td>
<td>53%</td>
<td>51%</td>
<td>69%</td>
<td>49%</td>
<td>51%</td>
</tr>
<tr>
<td>White (%)</td>
<td>33%</td>
<td>13%</td>
<td>4%</td>
<td>47%</td>
<td>90%</td>
<td>0%</td>
</tr>
<tr>
<td>African Amer (%)</td>
<td>65%</td>
<td>87%</td>
<td>90%</td>
<td>51%</td>
<td>1%</td>
<td>43%</td>
</tr>
<tr>
<td>Hispanic (%)</td>
<td>&lt;1%</td>
<td>0%</td>
<td>4%</td>
<td>0%</td>
<td>4%</td>
<td>58%</td>
</tr>
<tr>
<td>Other (%)</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>2%</td>
<td>5%</td>
<td>0%</td>
</tr>
</tbody>
</table>

*Intervention type classification captured based on classifications from Blueprints for Healthy Youth Development*
Larger sample size allows us to
- Study cross-over effects
- Have power to detect impacts on rare outcomes
- Explore subgroup analyses not previously possible

Rich measures collected across the life-course allows us to explore proximal prevention targets to understand mechanisms
Each dataset varies in the number of waves of data collected and timing of the waves.
Harmonizing data from disparate trials is challenging, particularly with complex study designs and both multilevel and longitudinal modeling.
PRE-ANALYTICAL HARMONIZATION

Data Discovery
- Obtaining codebooks
- Collect datasets

Pre-statistical Harmonization
- Create concordance table
- Merge and clean data, confirm response scales
- Identify common data domains

Statistical harmonization
- Harmonize response scales
- Create covariate informed factor scores
MEASUREMENT INVARIANCE

- Ensure the hypothesized factor being measured is the same under different conditions or across different groups
  - First, a baseline two-group model is established
  - Progressively specify more stringent constraints
  - Conduct a likelihood ratio difference test; If $p > .05$, MI is established
- Typically, when using multi-sample approaches, (a) discrete groups are (b) tested in succession
  - treatment vs. control, old vs. young, females, vs. males
  - Why not tested simultaneously?!
Extends the generalized factor analysis framework by allowing the four model parameters to vary as a function of covariates.
STEPS OF MNLFA

- Step 1: Graphical and descriptive analyses of items
- Step 2: Dimensionality Testing
- Step 3: Testing for factor and item differences
- Step 4: Estimate scale scores in pooled dataset
Step 1: Graphical and descriptive analyses of items

- Considering ordinal responses as binary
- Item "Yells at others" originally scored along a 6-point Likert scale

<table>
<thead>
<tr>
<th>Teacher-Yells at others</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almost Never</td>
<td>4,504</td>
<td>47.41</td>
<td>47.41</td>
</tr>
<tr>
<td>Rarely</td>
<td>2,080</td>
<td>21.89</td>
<td>69.31</td>
</tr>
<tr>
<td>Sometimes</td>
<td>1,634</td>
<td>17.20</td>
<td>86.51</td>
</tr>
<tr>
<td>Often</td>
<td>618</td>
<td>6.51</td>
<td>93.01</td>
</tr>
<tr>
<td>Very Often</td>
<td>421</td>
<td>4.43</td>
<td>97.44</td>
</tr>
<tr>
<td>Always</td>
<td>243</td>
<td>2.56</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>9,500</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teacher-Yells at others</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>4,504</td>
<td>47.41</td>
<td>47.41</td>
</tr>
<tr>
<td>Yes</td>
<td>4,996</td>
<td>52.59</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>9,500</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

© Musci et al. (2021)
Step 2: Dimensionality Testing

Conduct EFA to determine appropriate number of factors to retain
Tested 1, 2, 3, and 4 factors
Step 3: Testing for factor and item differences

© Musci et al. (2021)
## Final MNLFA Model Examining Covariate Effects on Factor Mean and Variance

<table>
<thead>
<tr>
<th>Covariate Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor Mean</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>-0.205</td>
<td>0.016</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.579</td>
<td>0.036</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Race</td>
<td>0.586</td>
<td>0.048</td>
<td>&lt; .001</td>
</tr>
<tr>
<td><strong>Factor Variance</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>0.107</td>
<td>0.018</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex</td>
<td>0.081</td>
<td>0.045</td>
<td>.072</td>
</tr>
<tr>
<td>Race</td>
<td>0.116</td>
<td>0.060</td>
<td>.052</td>
</tr>
</tbody>
</table>

*Note: SE = standard error.*
Testing for Factor and Item Differences

Examining DIF in Items Using a Sequential Modelling Building Approach

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariate</th>
<th># of Parameters</th>
<th>Log likelihood</th>
<th>Scaling factor</th>
<th>Factor Mean</th>
<th>Factor Variance</th>
<th>Item Intercept</th>
<th>Item Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Est</td>
<td>p</td>
<td>Est</td>
<td>p</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>27</td>
<td>-39166.368</td>
<td>1.0401</td>
<td>-0.358</td>
<td>&lt; .001</td>
<td>0.186</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.912</td>
<td>&lt; .001</td>
<td>0.341</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.778</td>
<td>&lt; .001</td>
<td>0.253</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X1</td>
<td></td>
<td>34</td>
<td>-38978.743</td>
<td>0.9970</td>
<td>-0.181</td>
<td>&lt; .001</td>
<td>0.051</td>
<td>&lt; .002</td>
</tr>
<tr>
<td>Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.476</td>
<td>&lt; .001</td>
<td>-0.047</td>
<td>&lt; .275</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.418</td>
<td>&lt; .001</td>
<td>0.117</td>
<td>&lt; .047</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td></td>
<td>33</td>
<td>-39086.742</td>
<td>1.0330</td>
<td>-0.357</td>
<td>&lt; .001</td>
<td>0.191</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.972</td>
<td>&lt; .001</td>
<td>0.356</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.754</td>
<td>&lt; .001</td>
<td>0.258</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td></td>
<td>33</td>
<td>-39121.400</td>
<td>1.0566</td>
<td>-0.364</td>
<td>&lt; .001</td>
<td>0.205</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.908</td>
<td>&lt; .001</td>
<td>0.320</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.786</td>
<td>&lt; .001</td>
<td>0.206</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td></td>
<td>33</td>
<td>-39132.380</td>
<td>1.0455</td>
<td>-0.368</td>
<td>&lt; .001</td>
<td>0.186</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.905</td>
<td>&lt; .001</td>
<td>0.351</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.820</td>
<td>&lt; .001</td>
<td>0.290</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Baseline model includes moderation on the factor mean and variance only (i.e., no item DIF). Moderation was investigated using three covariates, including study (an indicator variable for the study in which an individual is from), sex (0 = Female, 1 = Male, and race (0 = non-Black, 1 = Black).
Step 4: Estimate scale scores in pooled dataset

- Factor scores were estimating using
  - maximum likelihood estimation with robust standard errors and a logit link function.
- Estimated factor scores can then be used in subsequent analyses
Those with higher 'aggressive behavior' factor scores were significantly more likely to have attempted suicide (OR = 2.39, p = .001)

\[
\text{logit(Attempt Suicide)} = \beta_0 + \beta_1 \times \text{Male} + \beta_2 \times \text{Black} + \beta_3 \times \text{Intervention} + \beta_4 \times \text{Aggressive} + \beta_5 \times \text{Intervention} \times \text{Aggressive}
\]
LIMITATIONS & NEXT STEPS

- Inclusion of multi-level model—allowing the intervention to be at level 2, modeling intervention specific effects
- Inclusion of longitudinal modeling
- Improvement of the measurement model around exposure, baseline covariates, and outcomes
CHALLENGES TO IDA

- Principal Investigator participatory research
  - PI’s of the original studies should be involved at each step of the data harmonization process
- Just because the data is harmonized doesn’t mean its perfect, there are still limitations
- Sometimes there is too much data!
  - Work needs to be theoretically driven
MODEL FOR SUICIDE RISK (TURECKI & BRENT, 2016)
MEDIATORS & MODERATORS

- Aggressive & Disruptive behavior
- Deviant peer affiliation
- Substance misuse
- Sex
- Racial/ethnic group
- Sexual orientation

© Musci et al. (2020)
IMPLICATIONS

- Long-term cross over effects are important to estimate for early childhood prevention programming
- Advances in IDA methodology to account for heterogeneity in behavior and heterogeneity in program impact

© Musci et al. (2021)
ACKNOWLEDGMENTS

- Holly Wilcox, PhD (co-PI)
- Ryoko Susukida
- Masoumeh Aminesmaeili
- Katherine Masyn
- Joe Kush

- NIMH ---- MH122214
- Grant support for included datasets:
  - PIRC G1: MH42968, MH38725, MH40859, DA09897, DA004392
  - PIRC G2: MH38725, MH57005, DA11796
  - P2P: R305A080326
  - Fast Track Project: R18MH48043, R18MH50951, R18MH50952, R18MH50953, K05MH00797, and K05MH01027, DA016903, K05DA15226, RC1DA028248, P30DA023026, S184U30002
  - LIFT: MH054248
  - SAFEChildren: MH048248, R01DA020829
Innovations and Applications of Integrative Data Analysis (IDA) and Related Data Harmonization Procedures in Prevention Science

Co-editors:
- Antonio A. Morgan-López (RTI International),
- Rashelle J. Musci (Johns Hopkins University), &
- Catherine P. Bradshaw (University of Virginia & Johns Hopkins University)

LOI due September 1, 2021