Integrative (Big?) Data Analysis

Patrick J. Curran
Andrea M. Hussong
Daniel J. Bauer

University of North Carolina at Chapel Hill
Integrative Data Analysis

• The fitting of statistical models to raw data that have been pooled across multiple independent samples
• More of a methodological framework than a specific set of analyses
• Does the combination of multiple data sources provide a unique insight that is not possible in any individual data set when taken alone?
Integrative Data Analysis

• Potential advantages
  o efficient use of resources
  o greater age coverage
  o greater power
  o increased coverage of rare behaviors
  o greater sample heterogeneity
  o within- and across-study replication

• Potential disadvantages
  o massive data management; complex methods
  o critical: establish commensurate measures of all constructs across all contributing studies
Very Brief Example

Behaviorally-inhibited temperament

Negative affect, parent report

Externalizing, parent report

Peer substance use

Social withdrawal

Anxiety

Tension-reduction expectancies

Tension-reduction motives
Measurement: Depression

- Age
- Gender
- Ethnicity
- Alc. Diag.
- Study

Depression

- Lonely
- Cries
- Fears
- Worried
- Guilt
- Secretive
- Hopeless

Additional factors indicated by ellipses...
Scoring: Depression
Measurement: Substance Use

Diagram showing relationships between age, gender, ethnicity, Alc. Diag. Study, Poly-Substance Use, Alcohol-Specific Use, and various substance use behaviors such as frequency drunk, frequency 5+ drinks, frequency alcohol use, number of drinks at once, use of sedatives, use of marijuana, use of opiates, and use of stimulants.
Scoring: Substance Use

![Graph showing Substance Use over Age]
Applications of IDA to "Big Data"

• Depends on how define "big data"
• Work thus far focused on "long" data
  o concatenating data for additional cases
  o may be less relevant for big data
• Instead consider creating "wide" data
  o merging data for additional measures on cases
• But wide data much more challenging
  o uncertainty via probabilistic matching of cases
• Great promise, but much work needed