Approaches to Multi-level Modelling in Tobacco Control Research and Evaluation

Steve Brown
Department of Statistics and Actuarial Science
Co-Director, Population Health Research Group
University of Waterloo
Waterloo, Ontario

Interdisciplinary Capacity Enhancement
advancing the science to reduce tobacco use

Rehaussement des compétences par l'interdisciplinarité
faire progresser la science pour réduire l'usage du tabac

Multiple Influences on Health

- Understanding health behaviours or health outcomes can involve studying factors at:
- The subject level
- The family level
- The school/clinic/hospital/physician level
- The community level
- …

Multiple Influences on Health Outcomes

- Studying factors at different levels has implications for:
- Measurement
- Design
- Analysis

Multi-levels – Examples from School-based Research

- Levels could include students, (in) schools, (in) communities, (in) provinces, etc.
- As well, must consider the role of time – follow-up of original students and/or schools

e.g. Project Impact (Lovato et al)

- Aim is to understand the relationship between **adolescent** tobacco use and:
- Individual characteristics
- School-based programs & policies
- Tobacco control programs in the community (bylaws)
- Factors in the community environment

Project Impact - Sample

- Grades 10 & 11, municipalities with a population > 10,000
- 28 districts randomly sampled (BC-6, MB-3, ON-8, QC-8, NF-3)
- Within districts, 4 schools randomly selected
- Target was 112 schools, 12,320 students
- Follow-up of Impact schools 2 years later

Project Impact - Data Collection

- School Smoking Profile
- Northrup Survey on School Policies
- Survey of School Prevention/Cessation Programs
- Impac Teen School Neighbourhood Scan
- Municipal by-laws
- Municipal smoking rates
- Legislation
- Tobacco control spending
- SES, etc.

Multi-level Studies

- Data on the **response** variable is obtained from the **lowest level (a subject measure) – we won't aggregate**
- Data on explanatory variables available at **all** levels

Multi-levels – Examples from School-based Research

- Aim is to model the response as a function of explanatory variables at different levels
    - Student – level variables: age, sex, number of friends who smoke, …
    - School – level variables: programs, policies, distance from retail outlets, …
    - Community – level variables: by-laws, media, distance from contraband cigarettes, …
    - Provincial – level variables: existence of province-wide strategy,…
    - And interactions
    - Role of multiple measurements across time on different levels must also be considered

Multiple Influences

- Developing statistical models to handle data from multiple levels requires
  - Time
  - Patience
  - Decent software
  - Sense of adventure
- We will focus on some of the challenges in developing “Multi-level Models” using school-based research in tobacco control as an example

Multi-level Models

- Standard regression models assume independence between observations
- Model for student i:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + e_i,$$

x_i, w_i are covariates for subject i

β's are unknown parameters

e_i is a random variable that represents the combined effects of variables not modelled or controlled

Usual assumption: $e_i \sim N(0,\sigma^2)$ and independent

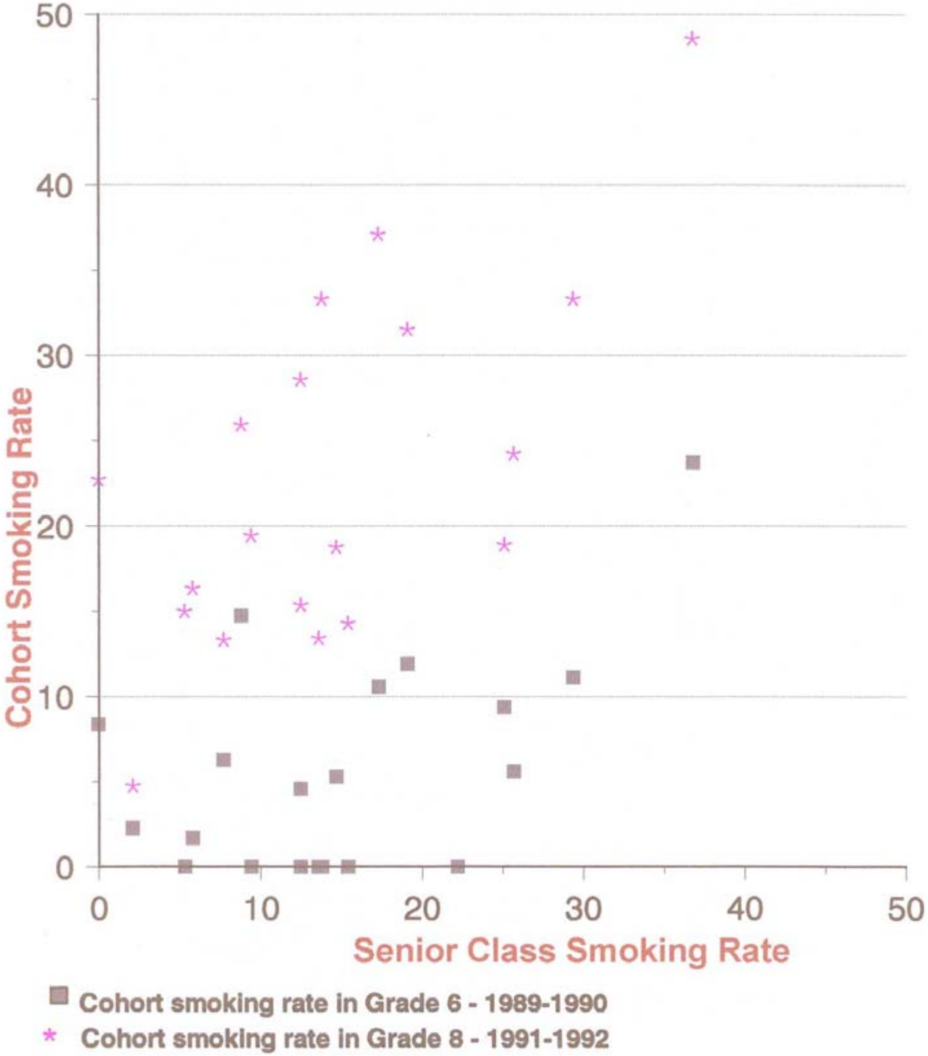
Multi-level Models

- Multi-level situations involve observations that are grouped in clusters
  - Students are clustered in schools (physical clustering)
  - Repeat observations on the same student are clustered within student (clustering in time)
  - Schools are clustered in communities (clustering in space)
- Clustering induces a correlation between observations in the same cluster
- Form of correlation is a modelling issue

Multi-Level Models

- Variation in smoking rates from school to school (cluster to cluster) is a fact of life
- Variation exceeds that which would be expected if children in the same schools were behaving independently (a la the binomial distribution)
- e.g. from WSPP3 (Cameron et al 1999)

Relation of School Risk* to Cohort Smoking Rates



* "School Risk" refers to the senior class (Grade 8) smoking rate measured when WSPP cohort students were in Grade 6.

Multi-Level Models

- Subjects in the same cluster are more similar than subjects in different clusters
- Correlation between students within schools leads to “extra-binomial” variation
- Ignoring leads to underestimating standard errors → calling estimates significant when not

Multi-level Models

- Need to account for the correlation between observations in the same cluster
- Two (common) ways to do this
    - Marginal Models (Population-Average Models)
    - Mixed Effects Models (Cluster-Specific Models)
- Important to understand the target of inference

Multi-Level Models

Type of Model	Target of Inference	Correlation	Procedure
Marginal Models	Population	Model for within-cluster association	GEE
Mixed Effects Models	Cluster (e.g. school)	Use random effects to induce correlation	Random effects models

Random Effects Models

- Illustration: Simple Structure – Simple Model
- k Students in n Schools
- Assume that the average value of the response will vary across schools – there is a school effect
- Need a parsimonious way to model this variability between schools
- Model with a random variable that represents the combined effects of variables that differ from school to school

Random Effects Models

- Model for student i in school (cluster) j

$$Y_{ij} = \beta_0 + b_{0j} + \beta_1 x_{ij} + \beta_2 w_{ij} + e_{ij},$$

where $b_{0j} \sim N(0,\sigma_s^2)$, and

$e_{ij} \sim N(0,\sigma^2)$ are assumed to be independent

- b_{0j} 's are the random effects – one per school
- b_{0j} 's represent the combined effects of those school-level variables that are not modelled or controlled
- We only estimate σ_s^2, so there is parsimony
- Adding b_{0j} 's induces a correlation between students in the same school
- x_{ij} and w_{ij} can be at individual or school level

Random Effects Models

- $Y_{ij} = (\beta_0 + b_{0j}) + \beta_1 x_{ij} + \beta_2 w_{ij} + e_{ij}$, illustrates a "random intercepts" model
- "Cluster-specific"
- Like fitting a collection of models with the same regression parameters (β's) but where the intercept varies from school-to-school
- If three levels then two random effects
- Ordering of the levels is important in specifying the model and making inferences

Random Effects Models

"Intercepts and slopes as outcomes models"

- Can also allow the slopes to vary randomly from school-to-school
- Like fitting a regression model within each school – get estimates of an intercept and a slope for each cluster
- Then could regress these estimates on cluster (school) - level variables

e.g. Project Impact (Lovato et al)

- Data from the first wave of Project Impact:
- Approximately 24,000 students from 81 schools
- Significant between-school variability in smoking rates after adjusting for individual and school level variables
- Variables associated with adolescent smoking:
  - **School-level** policy, and observed environment (e.g. students smoking near school) variables
  - **Student-level** policy perception, and perception of the environment variables

Some Design Examples - Basic Set-Up

- Assume 2 Conditions
  - Intervention and Comparison
  - Multiple clusters (e.g. Schools) per Condition
  - Multiple subjects per cluster (e.g. students per school)
- Assume 2 Time Points
  - Pre - and Post – Intervention
- (Possible to Randomize to Condition – often not feasible)

Basic Set Up

Intervention	
Pre	Post
School 1	School 1
School 2	School 2
School 3	School 3
School 4	School 4
School 5	School 5
School 6	School 6
⋮	⋮
School n	School n

Comparison	
Pre	Post
School 1	School 1
School 2	School 2
School 3	School 3
School 4	School 4
School 5	School 5
School 6	School 6
⋮	⋮
School n	School n

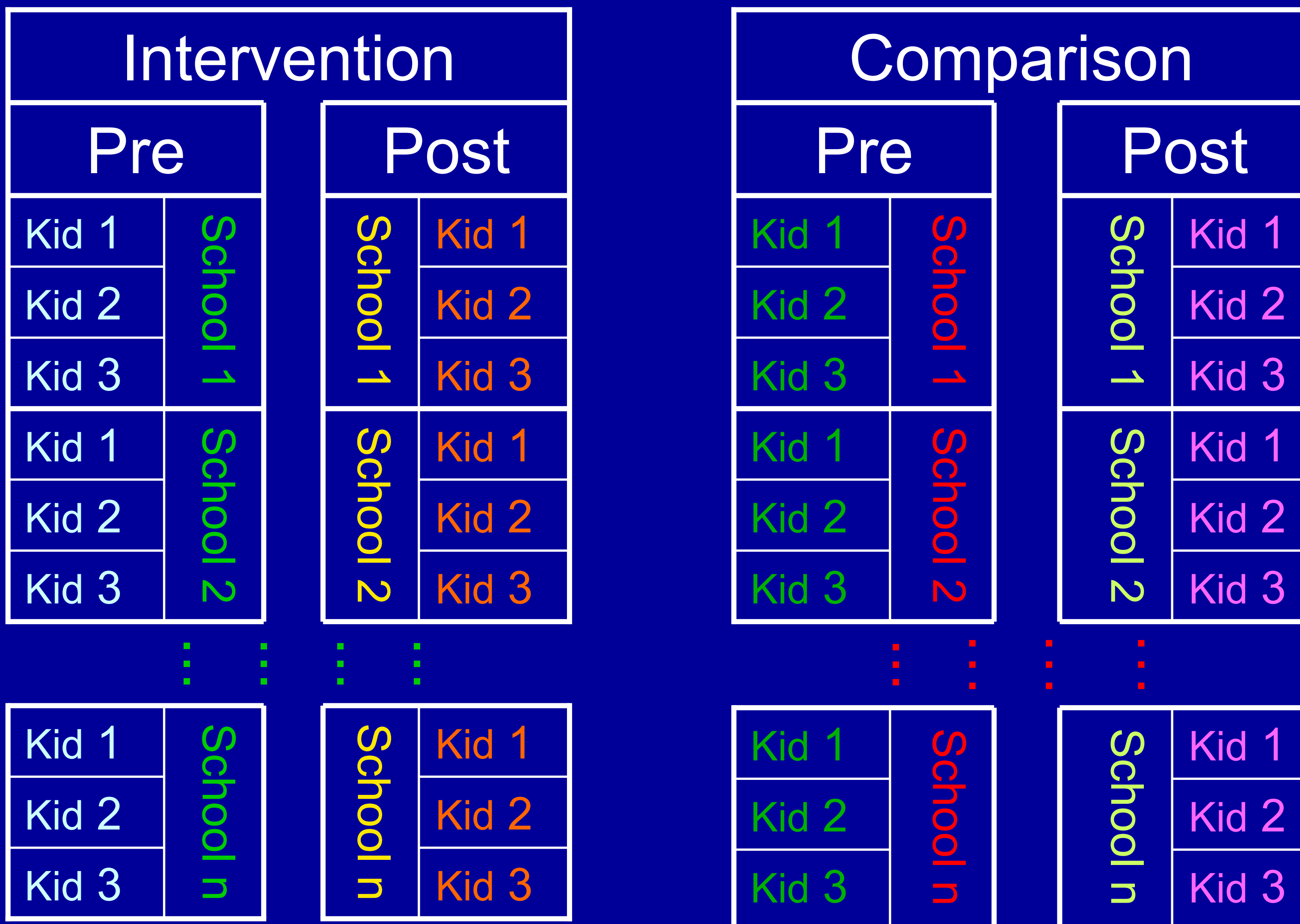
Some Design Choices

Design 1

- Different Schools at Pre- and Post-Test
- So ….. Different Students at Pre- and Post-Test
- E.g. Evaluation of Provincial Tobacco strategy with repeated cross-sectional surveys – comparison schools are from another province

Design 1

(Different Colours → Different Units)



Model for Design 1

Ordering

Kids in Schools in Times in Conditions

→

- Need to consider correlation between students in the same school
- Use school-level random effects to induce this correlation

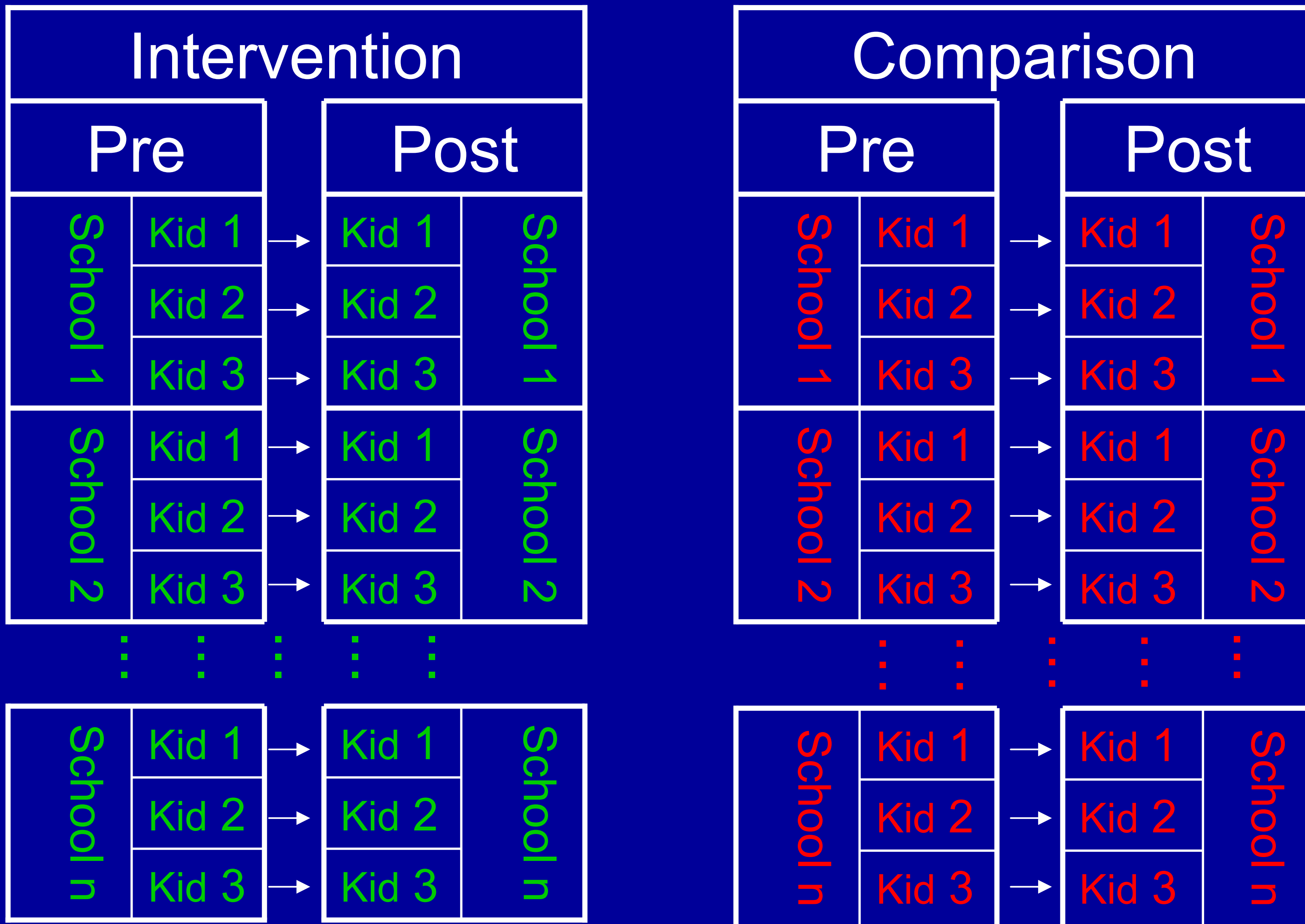
Some Design Choices

Design 2

- Same Schools at Pre- and Post-Test
- Students followed longitudinally from Pre-Post test
- Common randomized design – schools randomly assigned to intervention or comparison
- e.g. WSPP3 – (Cameron et al, 1999)

Design 2

Same colour → Same Unit



Model for Design 2

Ordering:

Times in Kids in Schools in Conditions

We need two types of random effects

- child-level random effects to induce correlation between measurements on the same child
- school-level random effects to induce correlation between students in the same school

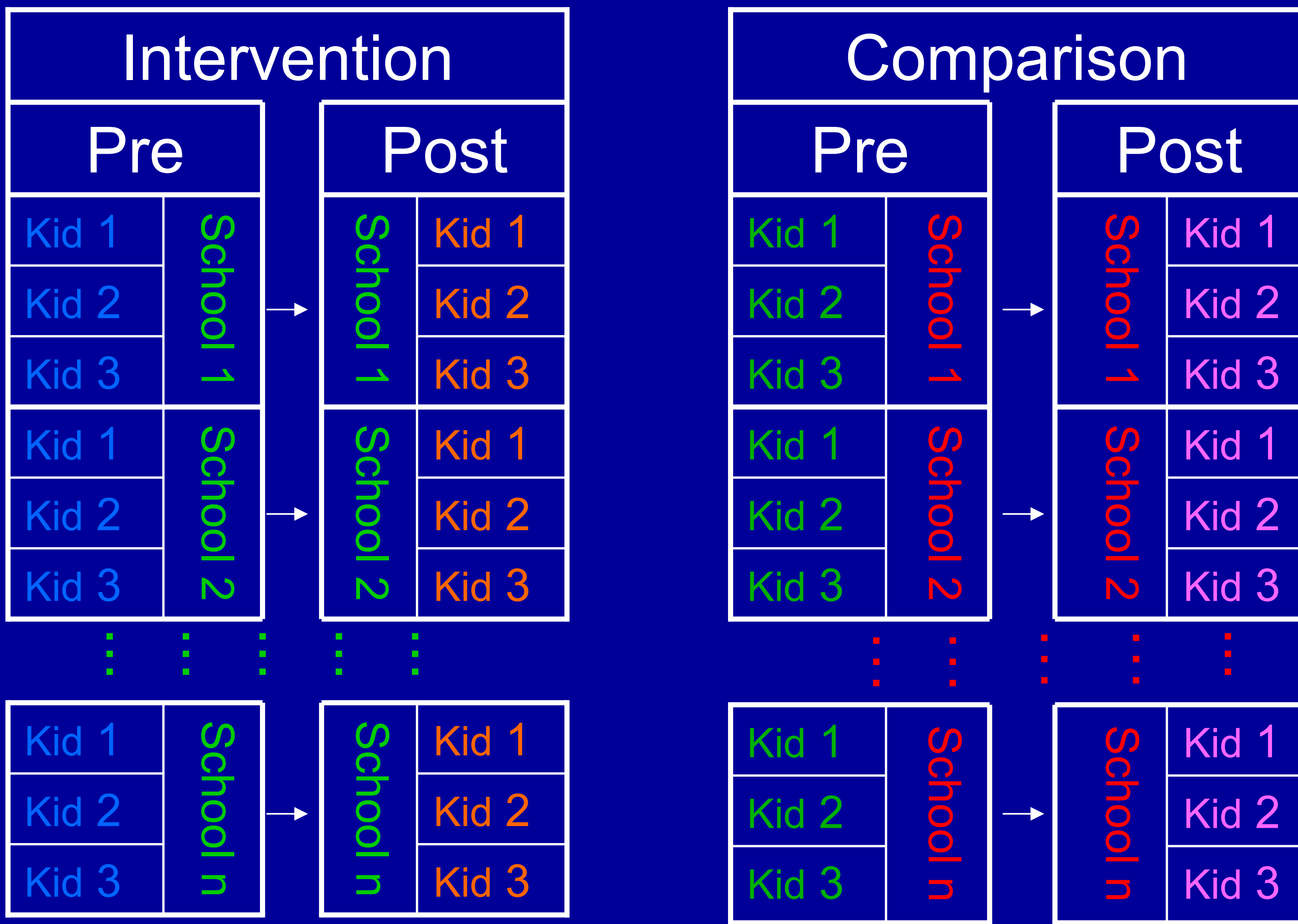
Some Design Choices

Design 3

- Same Schools at Pre- and Post-Test
- Different Students or (no linking of students) from Pre- to Post-Test
- e.g. Evaluation of Tobacco Control Policies in PEI (Murnighan, et al); Project Impact (Lovato, Brown, et al); WSPP4 – (Brown, Cameron, Manske et al)

Design 3

(Same colour → Same Unit)



Model for Design 3

Ordering:

Kids in Times in Schools in Conditions

→

- We need random effects for schools – the same effect at both times for a school
- We also need other random effects for schools at each time

Generalized Linear Models with Random Effects

- Extensions to logistic models are available
- e.g. logistic model with random effects:

$$Y_{ij} \sim \text{Binomial}(n_{ij}, p_{ij}), \text{ with}$$

$$\log(p_{ij}/(1-p_{ij})) = \beta_0 + b_{0j} + \beta_1 x_{ij} + \beta_2 w_{ij},$$

$$\text{where } b_{0j} \sim N(0, \sigma_s^2),$$

- Computations become more complicated
  - No closed form expressions
  - Involves numerical approximations

Sources of Explanatory Variables at Second level (Schools)

- Data from secondary sources
    - Statistics Canada data by postal code
- Data from administrative records
    - School size, % students bussed
- Data from school officials
    - Interviews, questionnaires
- Data from all students
    - Aggregated to level of school
- Data from selected students
    - E.g. older students to predict younger student behaviour

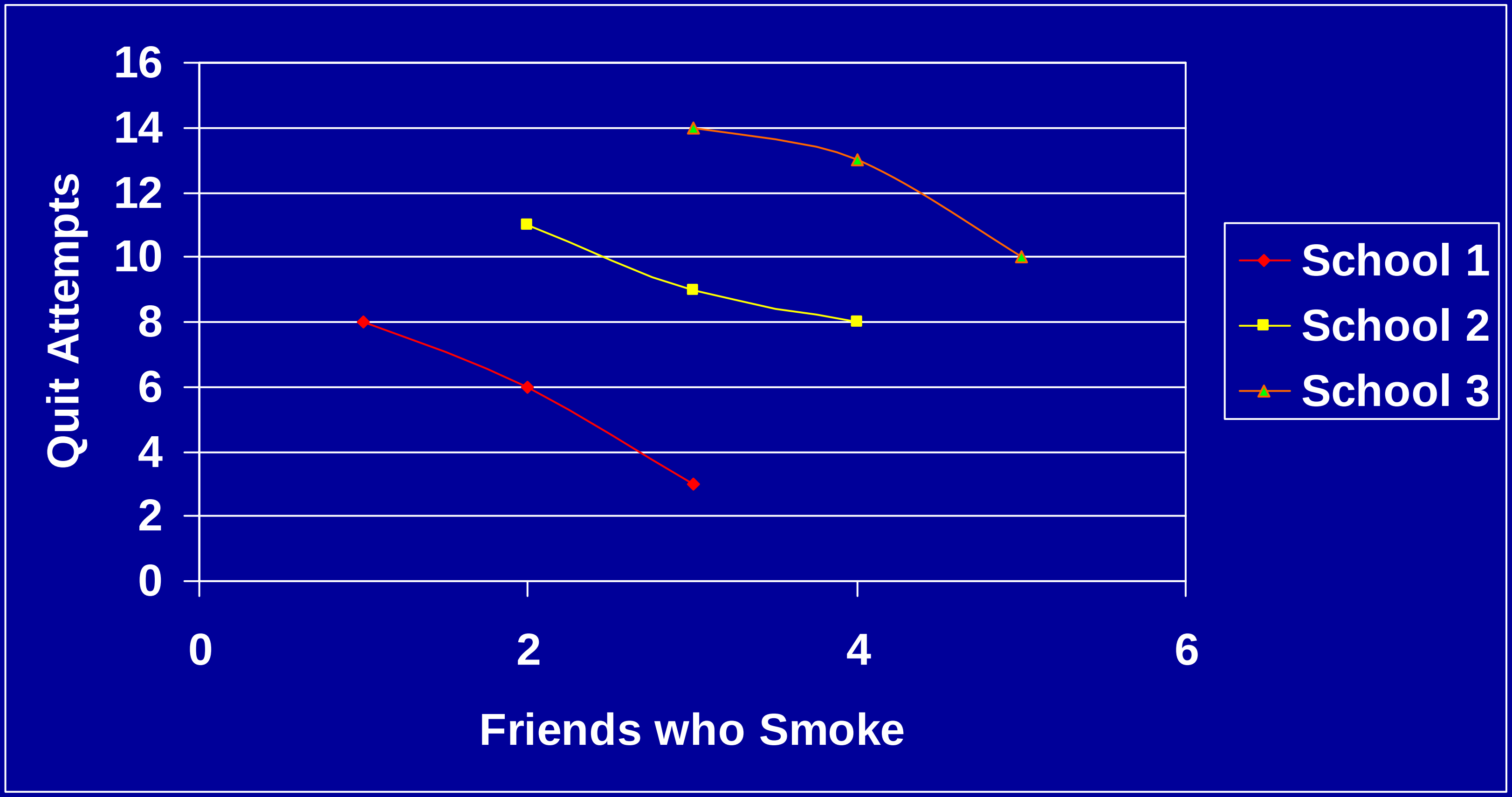
Higher Level Variables

- Objective measures at the level of school, community, etc. are problematic (e.g. “school environment”)
- Temptation to aggregate from level(s) below (i.e. average student responses to get measure for school)
- Does “Average perception = Reality”?
- What makes most sense theoretically (e.g. student perception of policy **or** actual policy?)

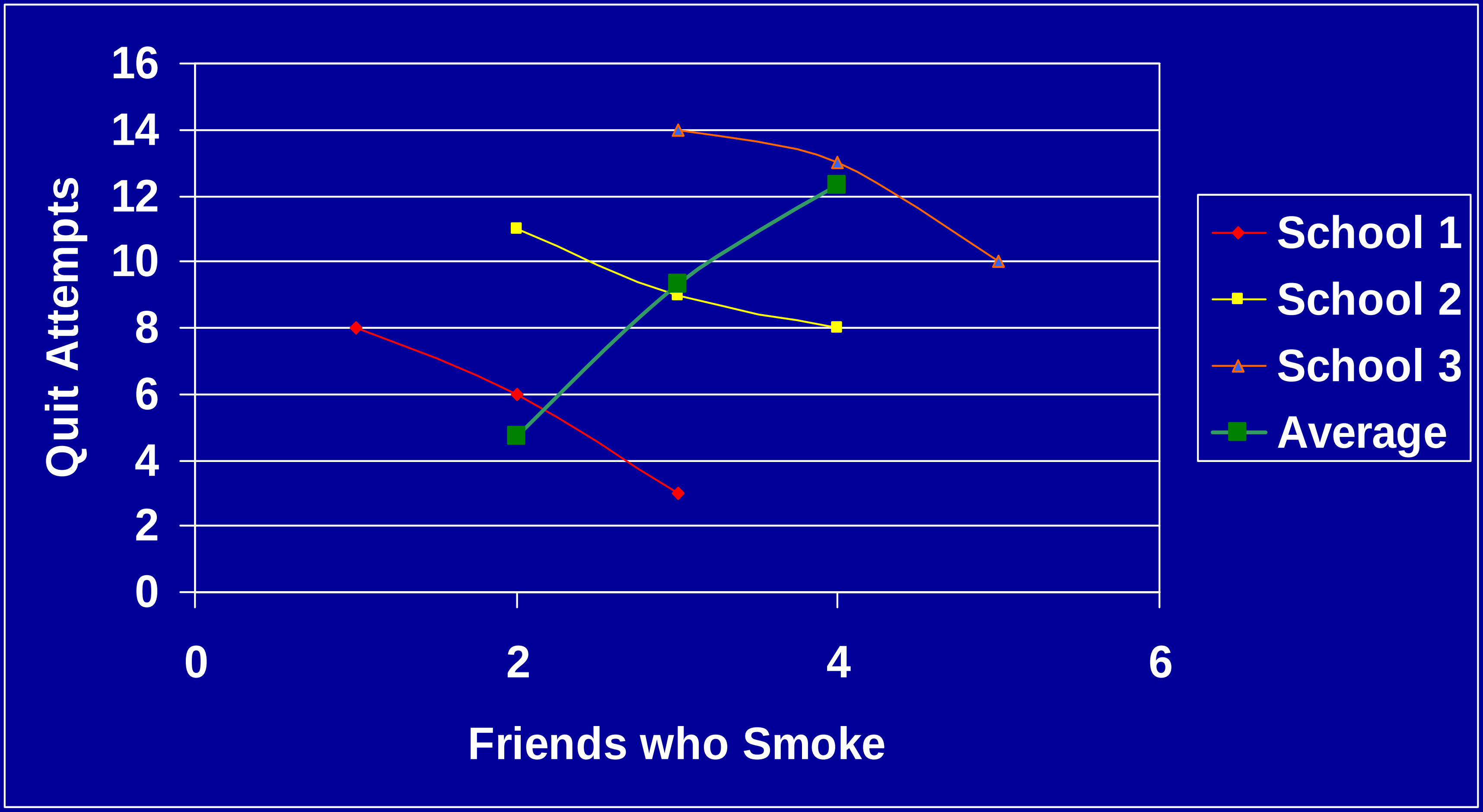
Higher Level Variables

- Could use average values from some students in schools as higher level variable (e.g. smoking rates in older students in school – WSPP3)
- Need to be careful in interpretation
- “Ecological fallacy”

Ecological Fallacy



Ecological Fallacy



Ecological Fallacy

Issues with Multi-level Models

- Measurement: Easy to imagine variables at higher levels – hard to operationalize in practice
- Power: More higher-level units more important than more lower-level units
- Fitting: How many levels are feasible? Software limitations (e.g. SAS – NLMIXED)

Issues with Multi-level Models

- Complicated Models: Interactions with variables at different levels may be important
- Extending Models: Desire to look at other outcomes (e.g. multi-state models, or bivariate outcomes)
- Model Reduction: Lots of variables! How to do variable selection in this context? By level? ….
- Missing Data: can occur at multiple levels, too

Conclusions

- Some form of multi-level modelling is necessary to handle correlation between observations within levels
- Measurement of higher level explanatory variables is tricky
- Often easier to imagine these models than to fit them
- Area of active research in (bio)statistics

Conclusions

- Multi-level Models: Panacea or Pandora's Box???

"All models are wrong. Some are useful"
(G.E.P. Box)