Latent variable and other methods for cohort and other cross-sample/cross measure comparisons.

Andrew Pickles

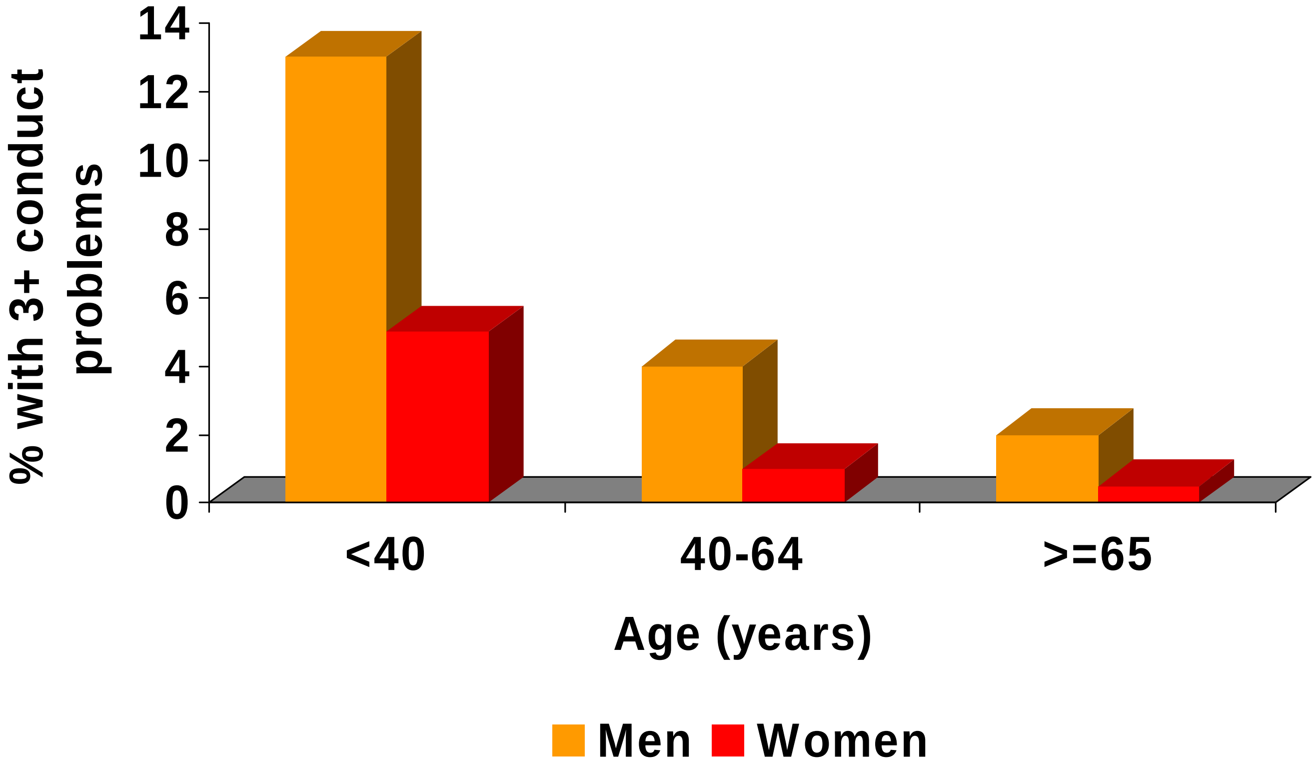
Biostatistics, Health Methodology Research Group

University of Manchester

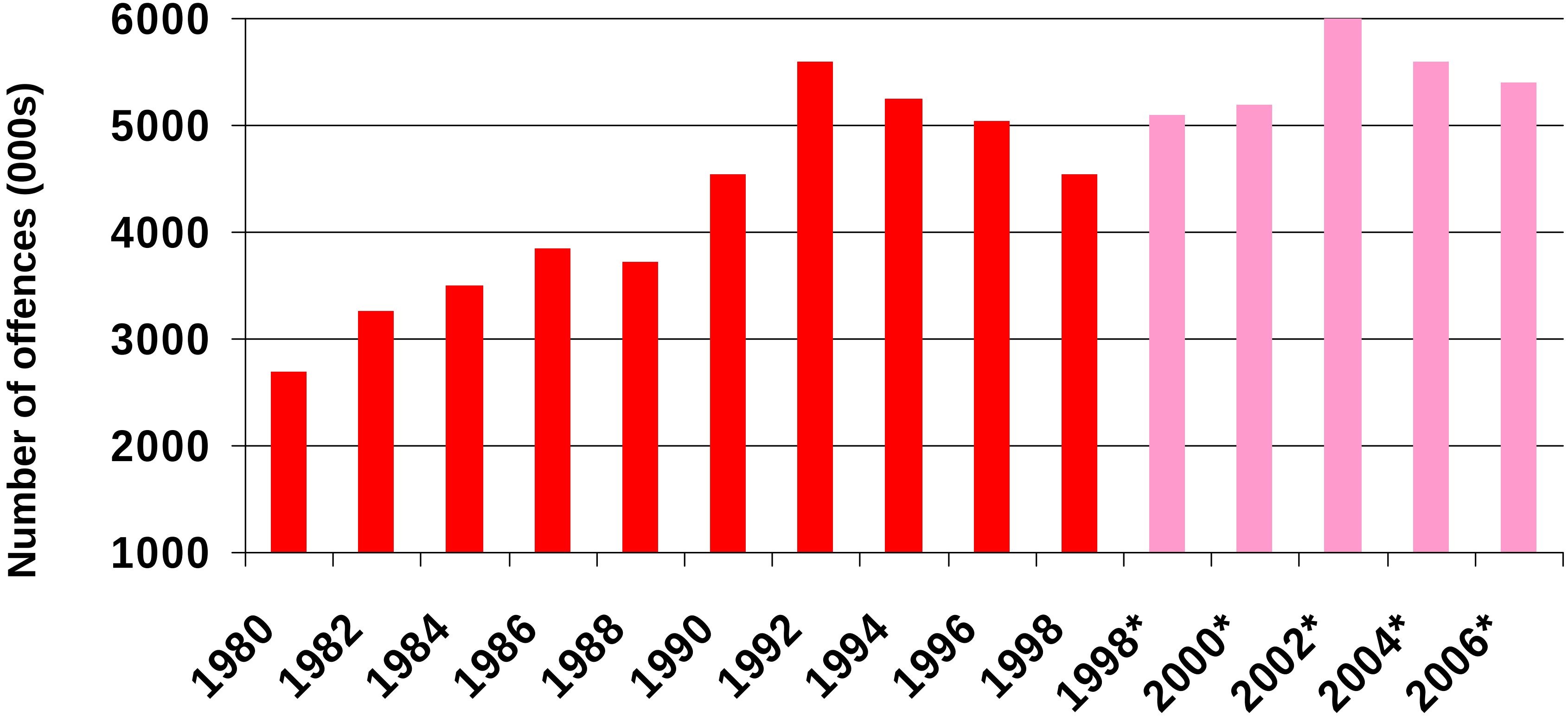
The Question

- How do we assess whether rates of some disorder or problem are increasing or decreasing?

3+ Conduct symptoms before age 15
Epidemiologic Catchment Area Study

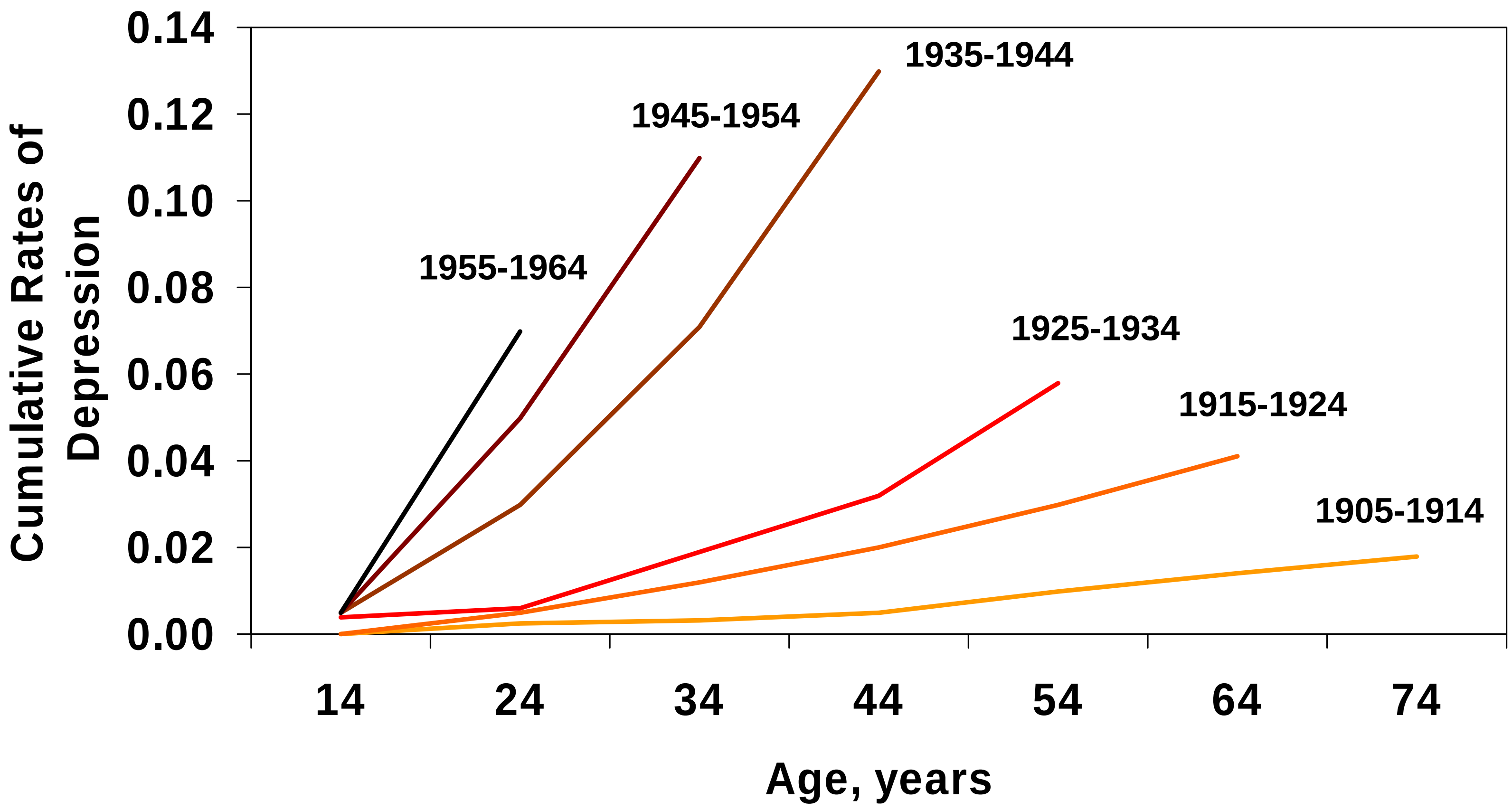


UK: Recorded crime 1980-2006



Life-time prevalence of major depression

Epidemiologic Catchment Area Study



Klerman & Weissman, 1989

Possible artefacts in reports

Retrospective report

- selective mortality / institutionalization
- effects of memory & recall
- changes in ‘psychological-mindedness’
- general reporting bias

Prospective Approach

- Changes in definitions, completeness and coverage of administrative recording
- Self, parent and teacher reports from national cohort studies

Parent ratings of adolescent mental health

Adolescent hyperactivity

- fidgeting
- inattention
- restlessness

Adolescent emotional problems

- misery
- worries
- fearful of new situations

Adolescent conduct problems

- fighting
- bullying
- stealing
- lying
- disobedience

Samples

	Cohort 1	Cohort 2	Cohort 3
Study	*National Child Development Study (NCDS)*	*1970 British Cohort Study (BCS70)*	*ONS Child Mental Health Survey (B-CAMHS99)*
Design	*Longitudinal: Birth-age 33*	*Longitudinal: Birth-age 29*	*Cross-sectional*
Parent ratings	*Rutter A*	*Rutter A*	*Goodman SDQ*
N	*10,499*	*7,293*	*868*
Age 16	*1974*	*1986*	*1999*

How do we map across Instruments? A calibration study

- Rutter and SDQ similar but not the same
- Calibration study: parents of additional samples of adolescents completed both the Rutter A scale and the SDQ.

N=219 parents from four secondary schools

N=78 parents of adolescents referred for psychiatric problems

N = 87 parents of a matched control sample.

The order in which parents completed the questionnaires was counterbalanced.

Calibration Study

- The calibration sample was used to impute Rutter questionnaire individual item scores, scale totals, and dichotomous problem scores for B-CAMHS99 study members on the basis of their SDQ scores.

- Multiple imputation was used, this having better statistical properties that become important where imputation error is larger

- Logit(Rutter item) = f(SDQ item

  $\sim= \beta_0 + \beta_1$ SDQ1+β_2SDQ2+…

  ordinal logit for Rutter scale or sub-scale

Calibration Task

Predicted items/subscales should possess all the variability and inter-item/scale associations as original A-scale items:

- problems of overfitting

  use Bayesian approach

  use pragmatic approach

For each measure of interest fit ordinal logistic regression to predict a Rutter-A output item/scale from a set of input SDQ predictors that consisted of:

- any closely matching input items
- relevant sub-scale scores
- overall scale score

Done separately for boys and girls

Multiple Imputation

While the ordinary‘fitted value’ from such a model may in some sense be the best predicted value for a particular individual, this value reflects neither the true variability in the population (fitted values have less variation than observed values) nor our level of uncertainty in that value.

Multiple imputation (Rubin, 1987) overcomes these problems

To reflect the uncertainty in our prediction equation, we first sampled the estimated coefficients of our ordinal logistic regressions by drawing values from a multivariate normal distribution defined by the estimated parameter covariance matrix.

$$\beta_{imp} = \sim N(\beta, \Sigma_{\beta})$$

Multiple Imputation

- We then used these to predict the probability of each feasible response value for each individual (e.g. 0, 1 or 2). One of these values was then picked with probability equal to this estimated response probability.
- These 3-steps were repeated 20 times to produce 20 B-CAMHS99 datasets with Rutter A scale measures.
- In this way the ‘made-up’ measures properly reflected behavioural variation as reported by the SDQ but the extent to which these datasets differed one from another properly reflected the uncertainty as to what value each of those ‘made-up’ values should be.

Multiple Imputation

In this application used own procedures but today would use Iterative chained equations as implemented in Patrick Royston's **`ice`** procedure

Van Buuren, S., H. C. Boshuizen and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18:681-694.

(Also see http://www.multiple-imputation.com)

Multiple imputation in Stata

Royston, P. 2004. Multiple imputation of missing values. Stata Journal 4: 227-241.

Royston, P. 2005. Multiple imputation of missing values: update. Stata Journal 5: 188-201.

Royston, P. 2005. Multiple imputation of missing values: update. Stata Journal 5: 527-536.

Royston, P. 2007. Multiple imputation of missing values: update. Stata Journal 7: 445-464.

Carlin, J. B., N. Li, P. Greenwood, and C. Coffey. 2003. Tools for analyzing multiple imputed datasets. Stata Journal 3: 226-244.

Typical ice code for sex specific imputing of Rutter domain scores from SDQ domain

```
ice rutemo ruthyp rutcd sdqemo sdqhyp sdqcd    /*
  */ sdqtot sex sdqemof sdqhypf sdqcdf sdqtotf /*
  */ using imputed.dta , m(20) /*
  */ passive(sdqemof:sdqemo*sex \ /*
  */         sdqhypf:sdqhyp*sex \ /*
  */         sdqcdf:sdqcd*sex \   /*
  */         sdqtotf:sdqtot*sex)  /*
  */ cmd(rutemo ruthyp rutcd sdqemo sdqhyp /*
  */ sdqcd sdqtot :ologit) cycles(30) replace
```

For calibration number of imputations m should be larger
The often recommended than 5 e.g. m(20) above.

Inference from Multiply Imputed samples

Analyse each of the 20 (m) datasets

Data are the same for

NCDS (Cohort 1) and BCS-70 (Cohort 2)

but may differ for the calibration imputed B-CAMHS99 (Cohort 3)

Parameter estimate =

mean of 20 estimates

Estimated parameter variance =

Estimated mean variance + ($1+m^{-1}$) estimated between dataset variance

- Often these calculations can be done automatically using the Stata commands “hotdeck”, “micombine” and “mim”

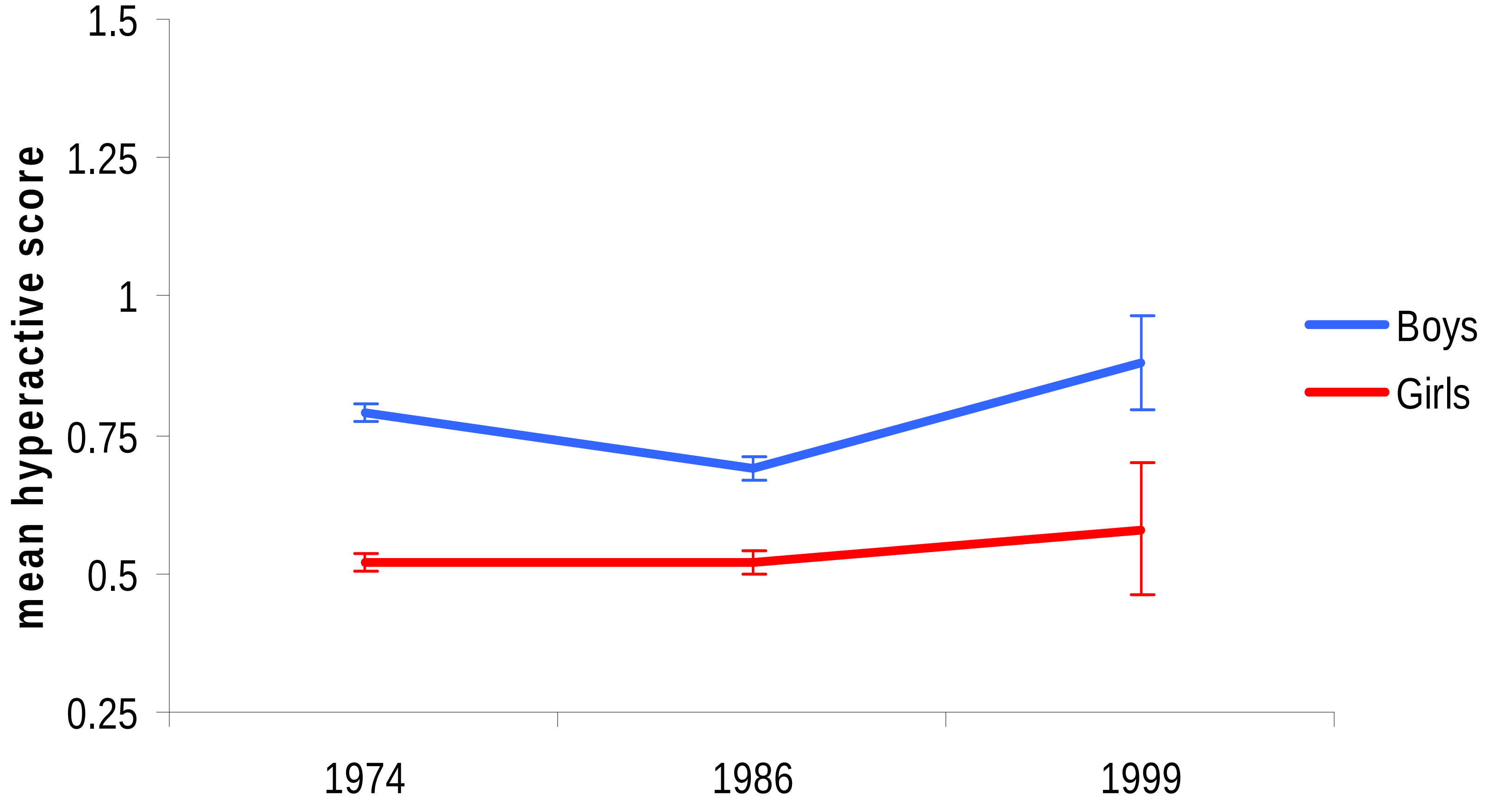
Typical analysis of multiply imputed datasets using `micombine`

Three cohorts with weights to adjust for sample design and attrition within each

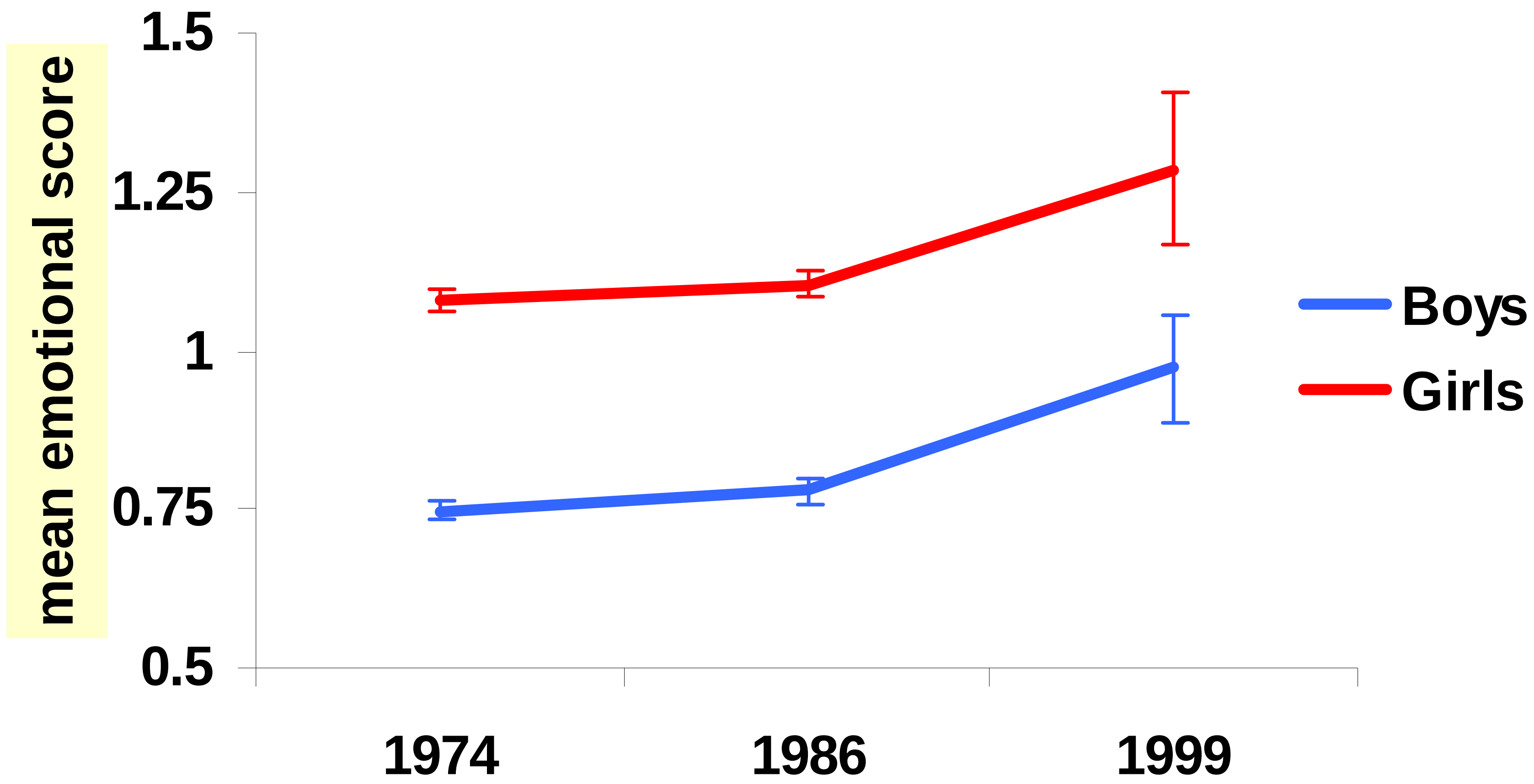
```
tab cohort, gen(c)
micombine ologit rutcd c2 c3 sex [pw=wt]
testparm c2 c3
recode rutcd (0/2=0) (3/max)=1
mim:logit rutcd c2 c3 sex [pw=wt]
mim:predict phat
mim:predict phatse, stdp
```

Results

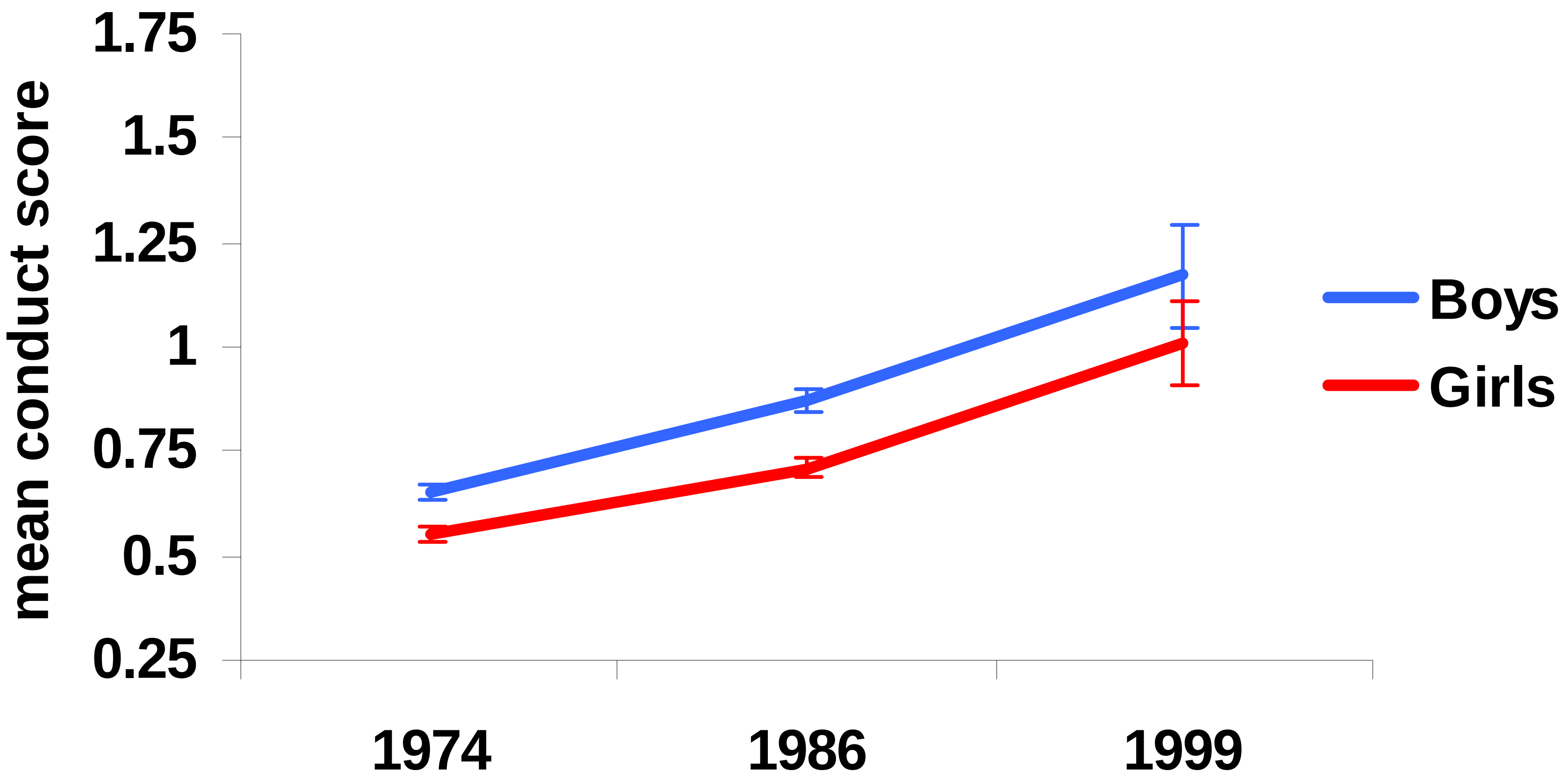
Time trends in adolescent hyperactive problems



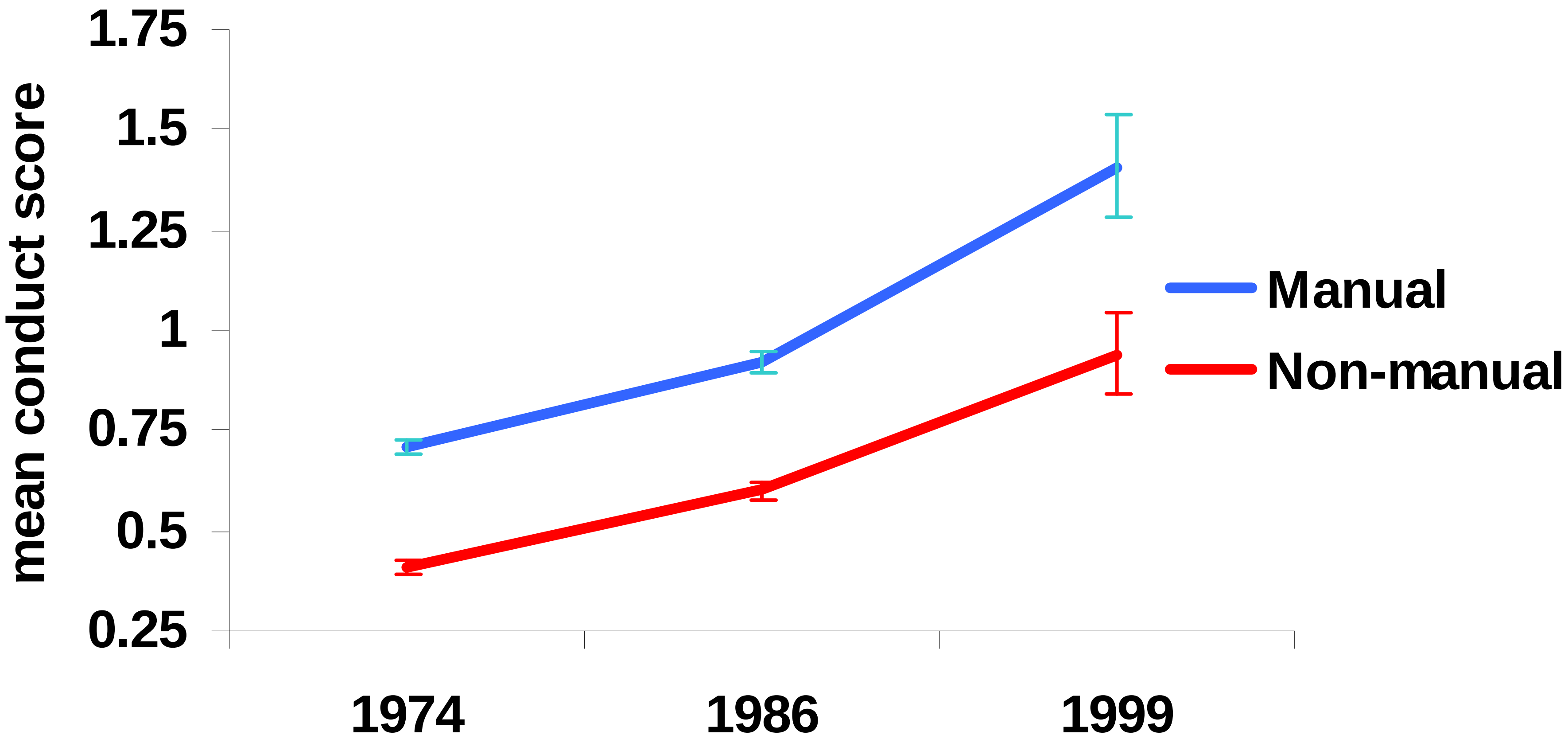
Time trends in adolescent emotional problems



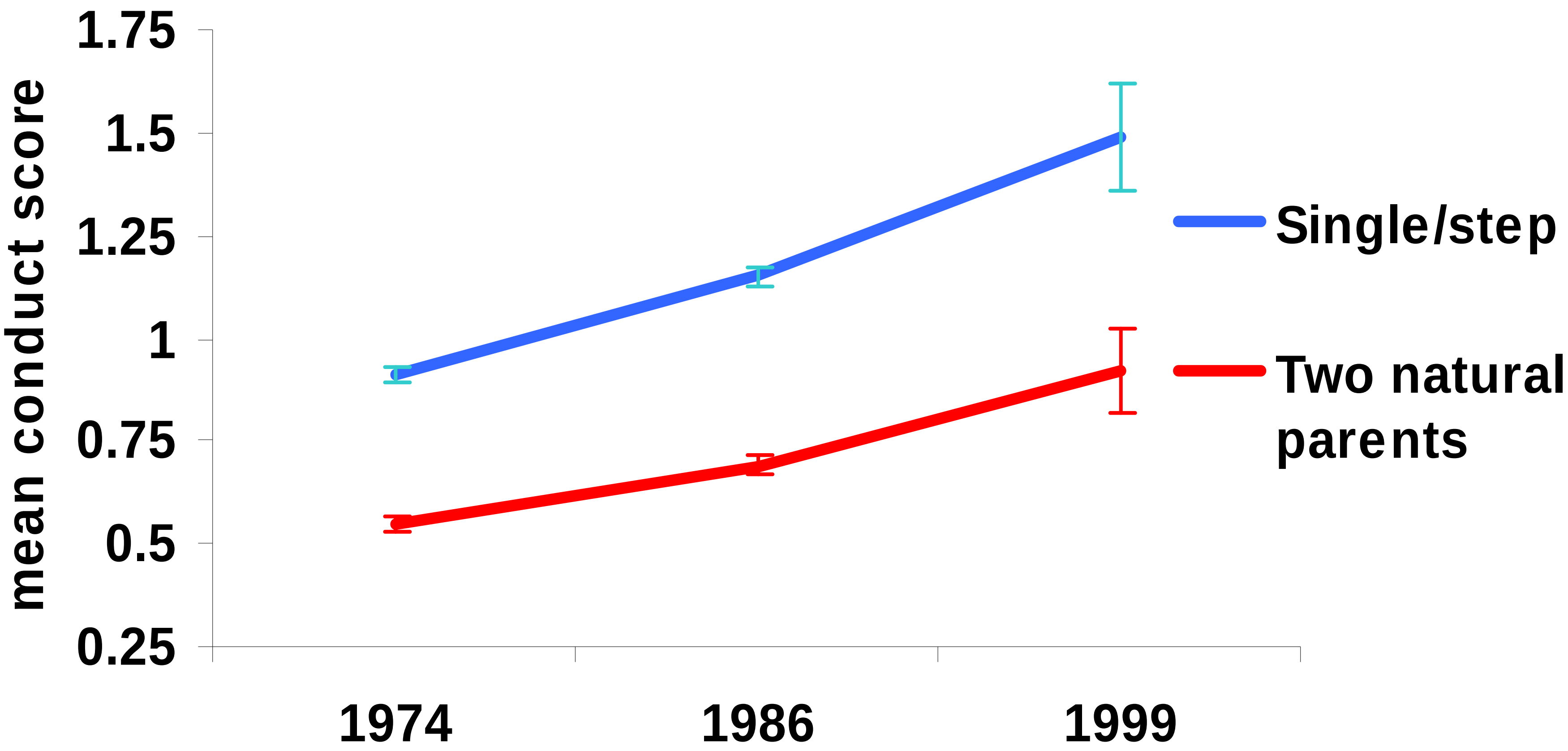
Time trends in adolescent conduct problems



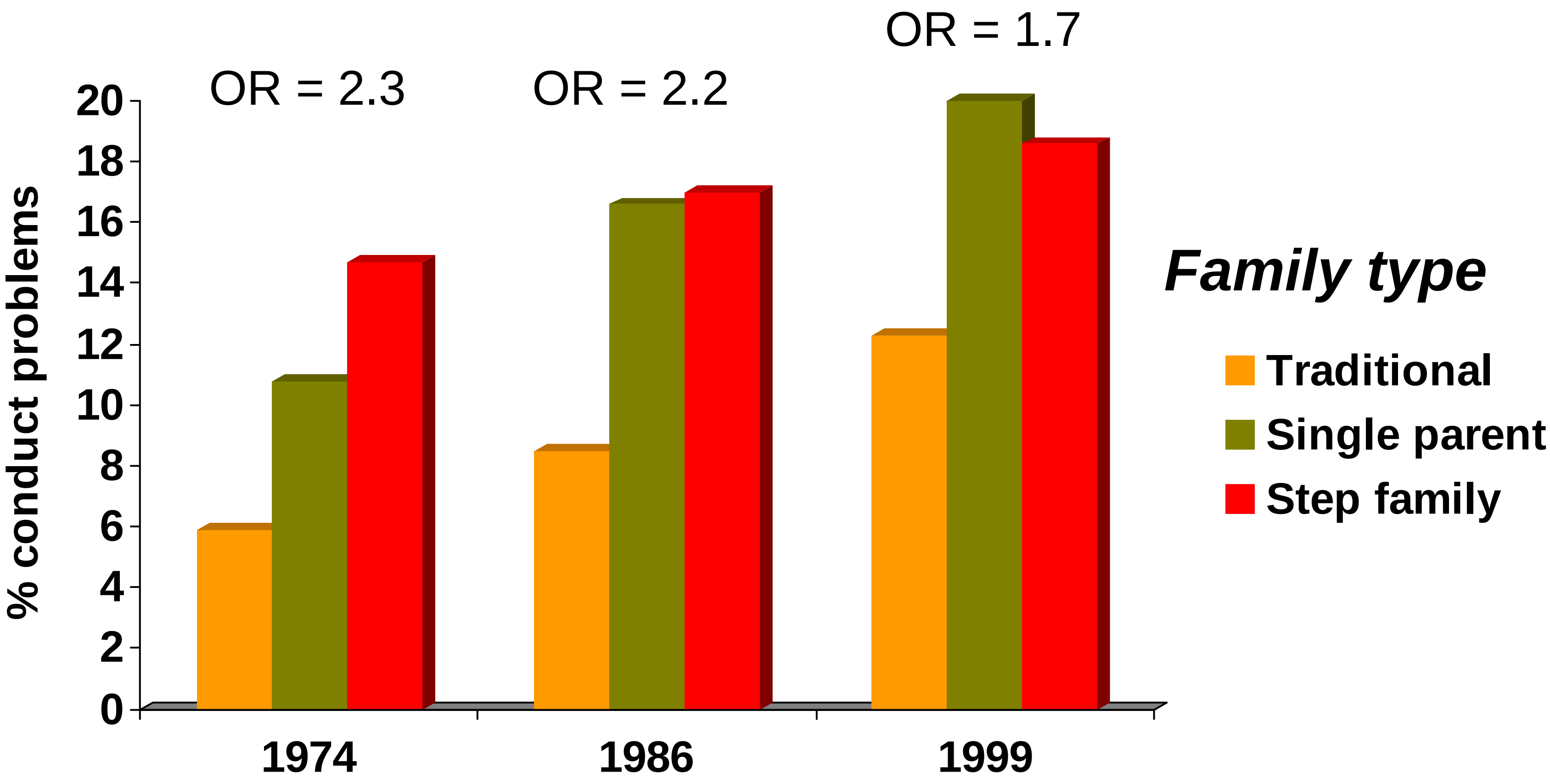
Trends in conduct problems: by social class



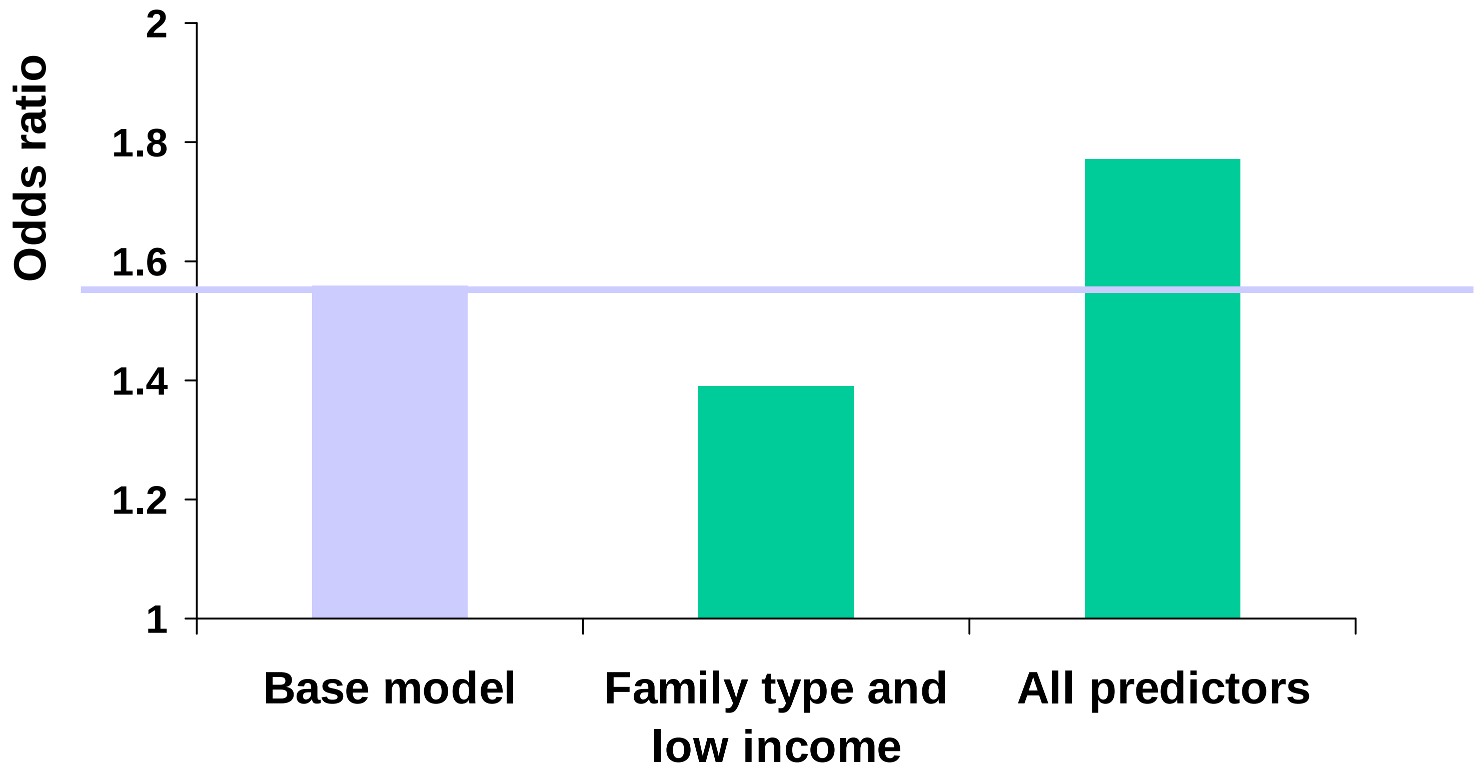
Trends in conduct problems: by family type



Adolescent conduct problems and family type



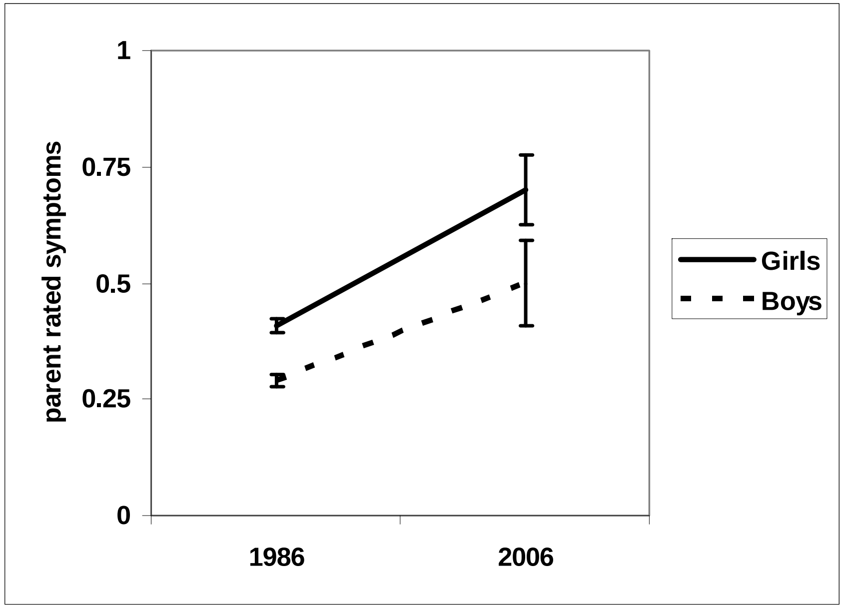
What residual time trend remains in conduct problems after "adjustment" for these other factors?

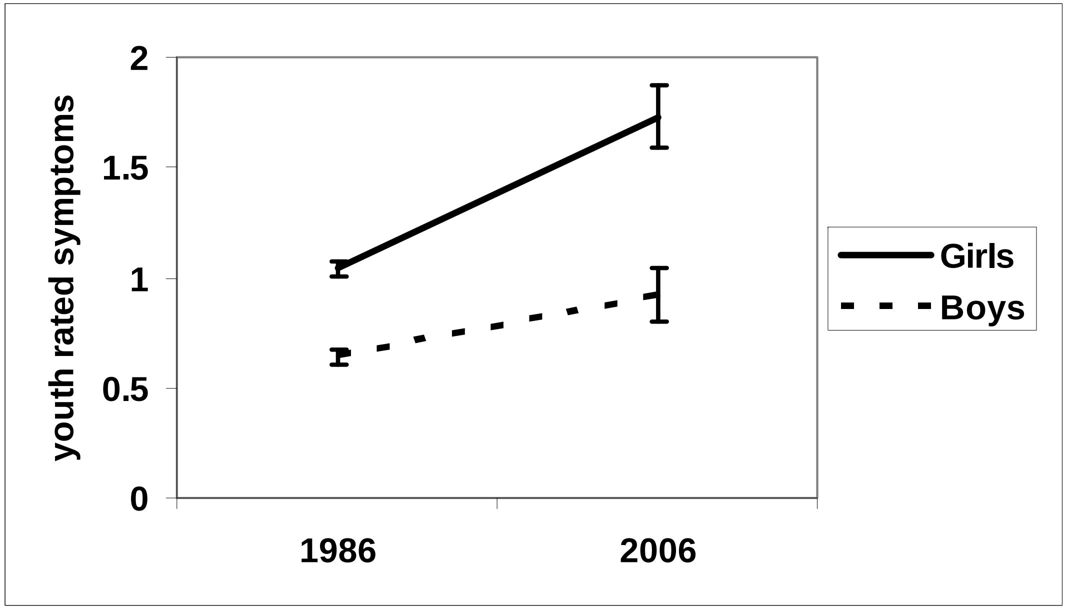


Is this an artefact of calibration? Replication without calibration

- Cohort 4: The 2002 and 2003 Health Surveys for England (Department of Health, 2003; National Statistics, 2004)
- 1401 children born 1st April 1988 to 31st March 1990 (mean age = 17.1 years, sd = 0.57 years)
- Surveyed in 2006 with same questions and scales as BCS 1970 birth cohort 1986 survey
- 715 adolescents and 737 parents (86% mothers, 14% fathers) responded to the 2006 survey
- Weighting to make comparable to general population

Trends based on identical questions





Parent

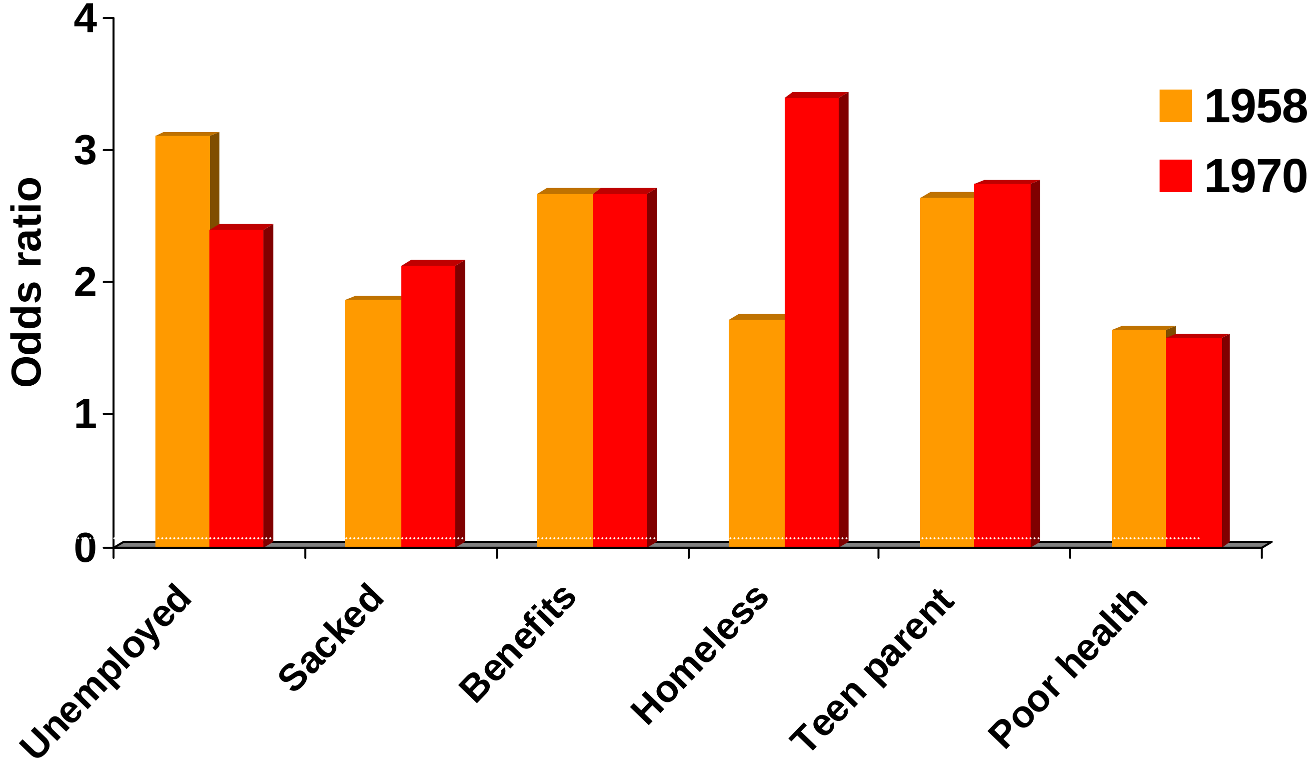
Self Report

Are the parents rating the same thing in each cohort?

Can we validate against adult outcomes of conduct problems?

- Socio-economic problems
- Relationship difficulties
- Teenage parenthood
- Poor physical health
- Poor mental health
- Crime
- Multiple difficulties

Conduct problems:
Age 30 outcomes 1958 & 1970 cohorts



Collaborators, references, funder and future findings

Collishaw, S., Maughan, B., Goodman, R. & Pickles, A. (2004) Time trends in adolescent mental health. *Journal of Child Psychology and Psychiatry. 45, 1350-1362.*

Goodman, R. Iervolino, A., Collishaw, S., Pickles, A. & Maughan, B. (2007) Seemingly minor changes to a questionnaire can make a big difference to mean scores: a cautionary tale. *Social Psychiatry and Psychiatric Epidemiology. 42(4):322-327*

Collishaw, S., Goodman, R., Pickles, A. & Maughan, B. (In Press) Modelling the contribution of changes in family life to time trends in adolescent conduct problems. *Social Science & Medicine*

Funded by Nuffield Foundation

Analyses of younger cohorts suggest this upward trend may have turned…

Latent variables as implicit imputed values

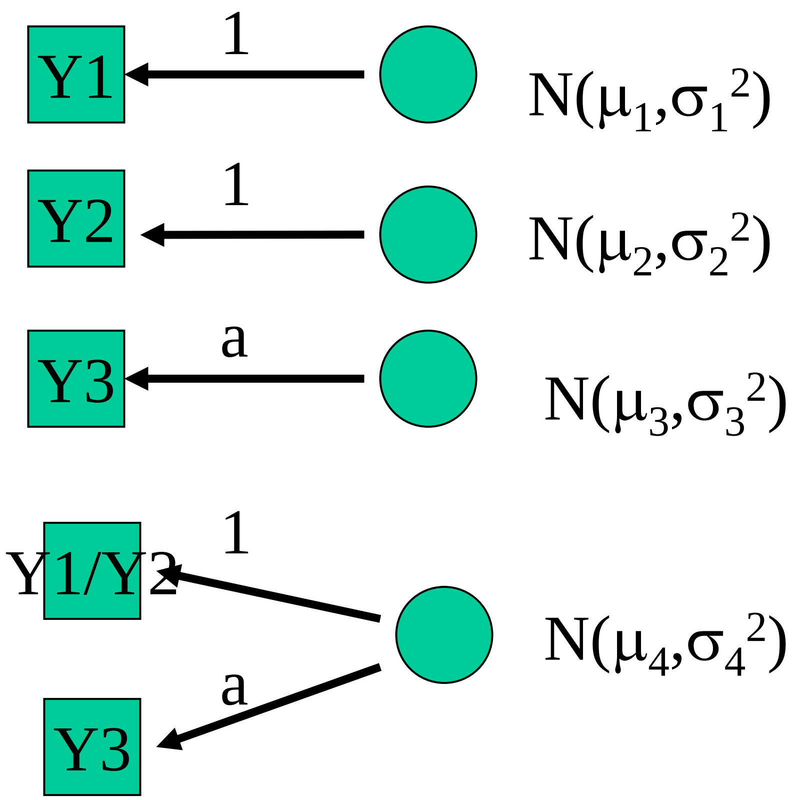
- Structural equation models often make use of factors, which are typically not directly observed variables

- We can estimate the “effects” of factors without actually estimating the factor scores. In a model the factors are defined by the model structure and the identifiability of the relevant parameters.

- Provided we can achieve identifiability with plausible restrictions/ assumptions we can construct models where missing variables are represented by factors or latent variables

SEM for Adolescent Cohort Comparisons:

where measures are different for 1958, 1970 and CAMHS-99



1958 cohort Y1

1970 cohort Y2

1999/2004 cohorts Y3/Y4

Calibration
Sample
(Y1,Z)

Estimate ordinal response latent variable model in gllamm (or Mplus!). Exploits MAR property of maximum likelihood.

Internal Calibration Samples: Diet and CHD

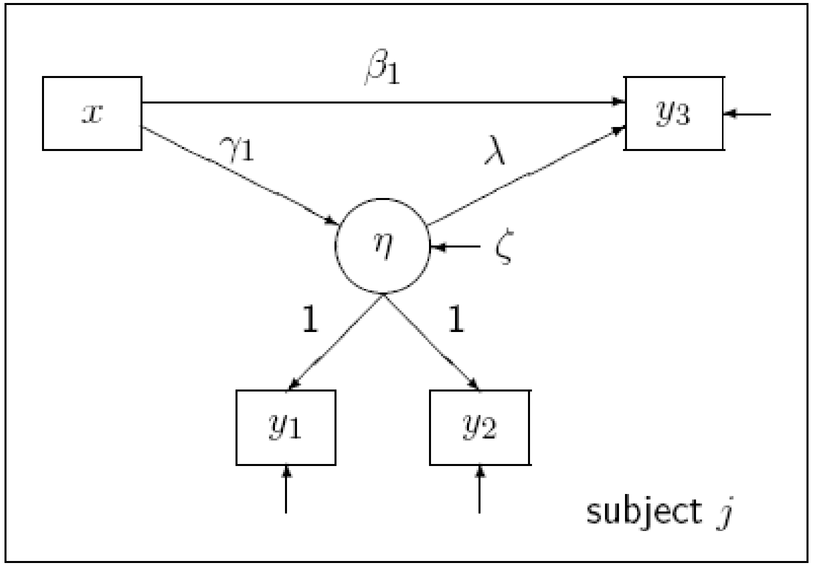
- Data from Morris et al (1977) analysed by Rabe-Hesketh, Pickles & Skrondal (2003)

- 337 middle-aged mean 7-day weighed food intake, and followedup for CHD events

- 77 repeated food intake record 6-months later

  - an *internal* calibration sample accounting for measurement error.

Accounting for error in covariate measurement



Note:

- including x_j both in disease and exposure models allows for a *direct effect* of x_j on disease (β_1) as well as an *indirect effect* through true exposure $(\lambda\gamma_1)$.

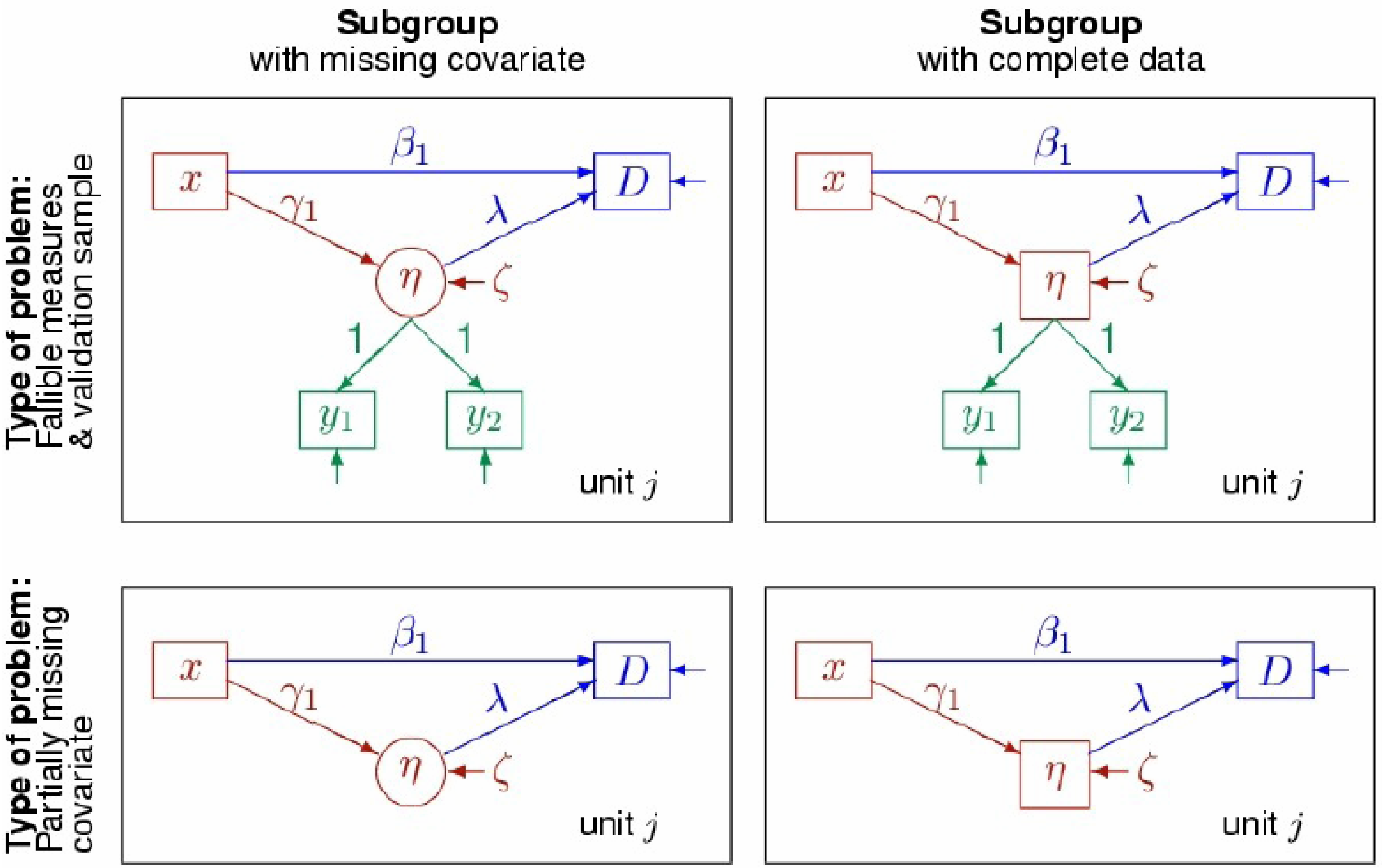
- If there is no direct effect $(\beta_1 = 0)$, $\delta_1 = \gamma_1\lambda$, the product of the effect γ_1 of x_j on y_{ij} and the effect λ of η_j on y_{3j}, a nonlinear constraint!

gllamm model for mixed type response

id	resp	diet	chd	var	bus_chd	bus_diet
217	3.06	1	0	1	0	0
217	0	0	1	2	0	0
218	3.14	1	0	1	0	0
218	0	0	1	2	0	0
219	2.75	1	0	1	0	0
219	2.7	1	0	1	0	0
219	0	0	1	2	0	0

```
gllamm resp chd bus_chd diet bus_diet, i(id) /*
*/ eqs(load) nocons lv(var) fv(var) /*
*/  link(iden logit) fam(iden bin) adapt
```

Further types of problem: (1) gold-standard (2) missing covariate

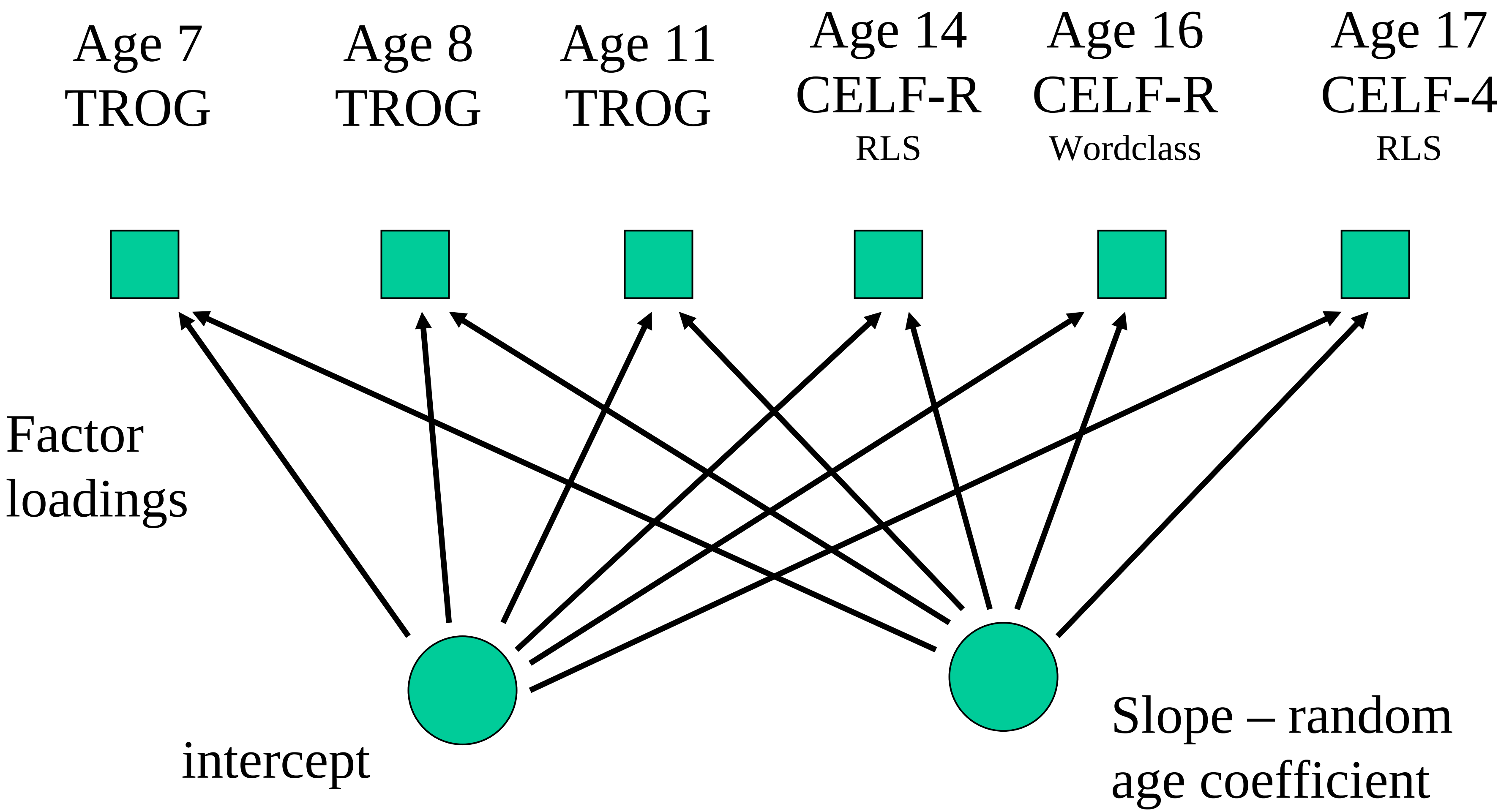


Developmental Trajectories in Specific Language Impairment (SLI)

SLI

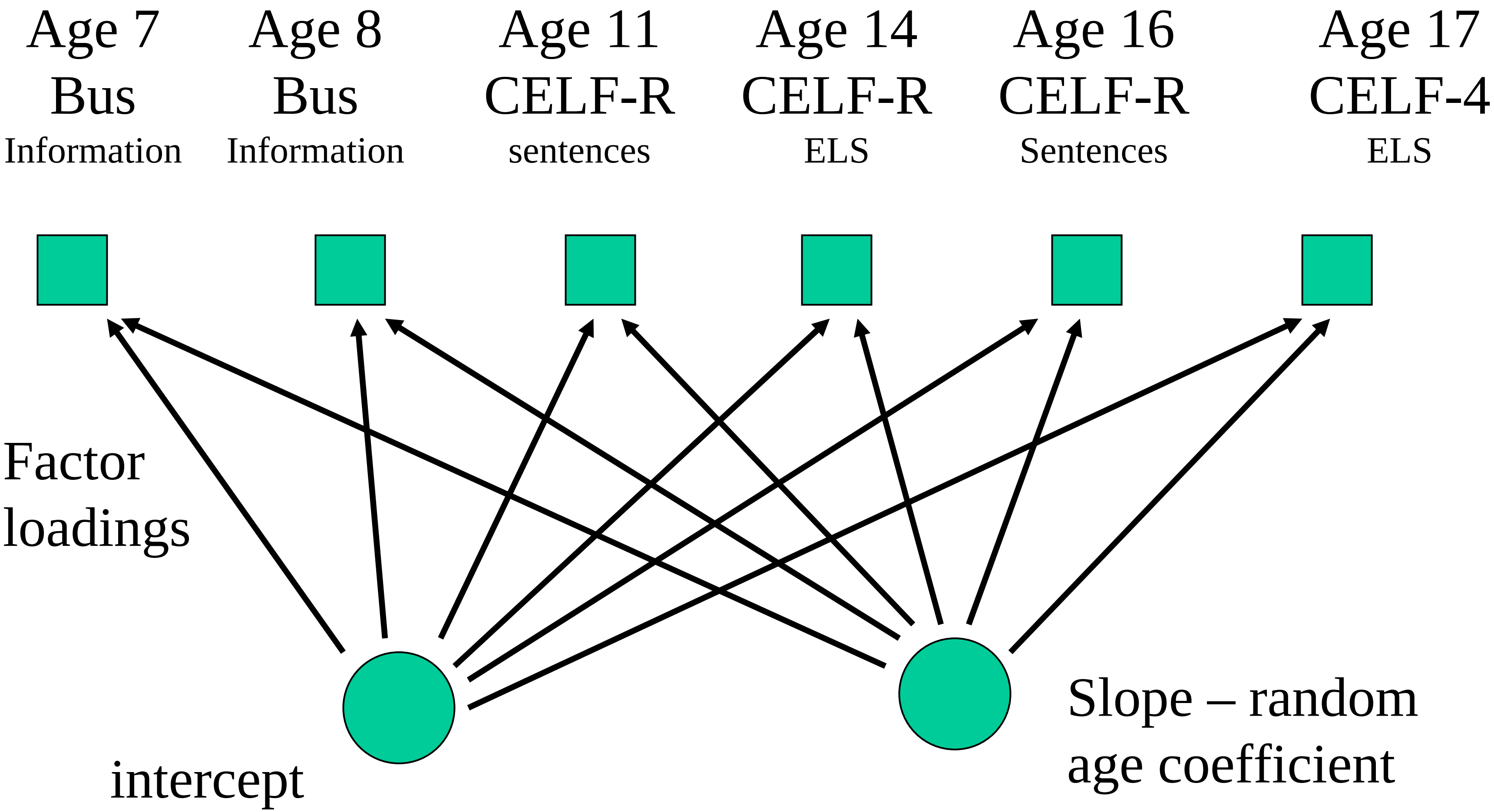
- SLI is a heterogeneous disorder with a variety of language and related problems
- A number of studies have focused on outcomes, however few have examined developmental language growth patterns and how this may inform the classification (subgrouping) of SLI
- Heterogeneous nature may lead to different developmental trajectories with differing associated symptomatology
- Manchester Language Study – cohort of children in special language schools followed from age 6.

Receptive Language-scaled growth curve



Impose factor loading and error variance constraints where the same measure is being used at different occasions

Expressive Language –"scaled" growth curve



Impose factor loading and error variance constraints where the same measure is being used at different occasions

Random coefficient models in GLLAMM

- One covariate multiplies each latent variable,

$$\eta_m^{(l)} z_{m1}^{(l)} \quad (\lambda_{m1}^{(l)} = 1)$$

- e.g. Latent growth curve model for individuals j (level 2) observed at times t_{ij}, $i = 1, \cdots, n_j$ (level 1)

$$\text{Linear predictor: } \nu_{ij} = \beta_1 + \beta_2 t_{ij} + \eta_{1j}^{(2)} + \eta_{2j}^{(2)} t_{ij}$$

β_1, β_2 :	mean intercept and slope
$\eta_{1j}^{(2)}$, $\eta_{2j}^{(2)}$:	random deviations of unit-specific intercepts and slopes from their means

Generalized random coeff. model in GLLAMM

$$\nu = \mathbf{x}'\boldsymbol{\beta} + \sum_{l=2}^{L}\sum_{m=1}^{M_l} \eta_m^{(l)} \mathbf{z}_m^{(l)\prime} \boldsymbol{\lambda}_m^{(l)}$$

For identification, $\lambda_{m1}^{(l)} = 1$

- Fixed part: $\mathbf{x}'\boldsymbol{\beta}$ as usual
- Random part:
  - $\eta_m^{(l)}$ is mth latent variable at level l, $m = 1, \cdots, M_l$, $l = 2, \cdots, L$
    Can be a factor or a random coefficient
  - $\mathbf{z}_m^{(l)}$ are variables and $\boldsymbol{\lambda}_m^{(l)}$ are parameters
  - Unless regressions for the latent variables are specified, latent variables at different levels are independent whereas latent variables at the same level may be dependent

Discrete latent variables in GLLAMM

- Linear predictor in two-level models:

$$\nu_{ij} \;=\; \mathbf{x}'_{ij}\boldsymbol{\beta} + \sum_{m=1}^{M} \eta_{jm}\mathbf{z}'_{mij}\boldsymbol{\lambda}_m, \qquad \lambda_{m1} = 1$$

- Latent variable vector η_j for unit j with discrete values (or locations) $e_c, c=1,\cdots,C$ in M dimensions
- Units in same latent class share the same value or location e_c
- Probability that unit j is in latent class c is $\pi_c = \frac{\exp(\varrho)}{1+\exp(\varrho)}$
- Two parameterizations:

1. **non-centered**: e_c, C locations freely estimated
2. **centered**: $\tilde{e}_c$, $C-1$ locations estimated, last location determined by constraint $\sum_c \pi_c \tilde{e}_c = 0$
   Allows mean structure to be modeled using $\mathbf{x}'_{ij}\boldsymbol{\beta}$

Language Model Specification

- Discrete trajectory classes located in 4 dimensions (2 intercept x 2 slope)
  - allows random effects to be correlated across expressive and receptive.
  - Increase number of classes and select "best-fit" model

- 6 receptive measures using 4 tests
  - 4 intercept factor loadings constrained equal to corresponding slope factor loadings
  - 4 measurement error variances
- 6 expressive measures using 4 tests (2 near parallel)
  - 3 intercept factor loadings constrained equal to corresponding slope factor loadings
  - 3 measurement error variances

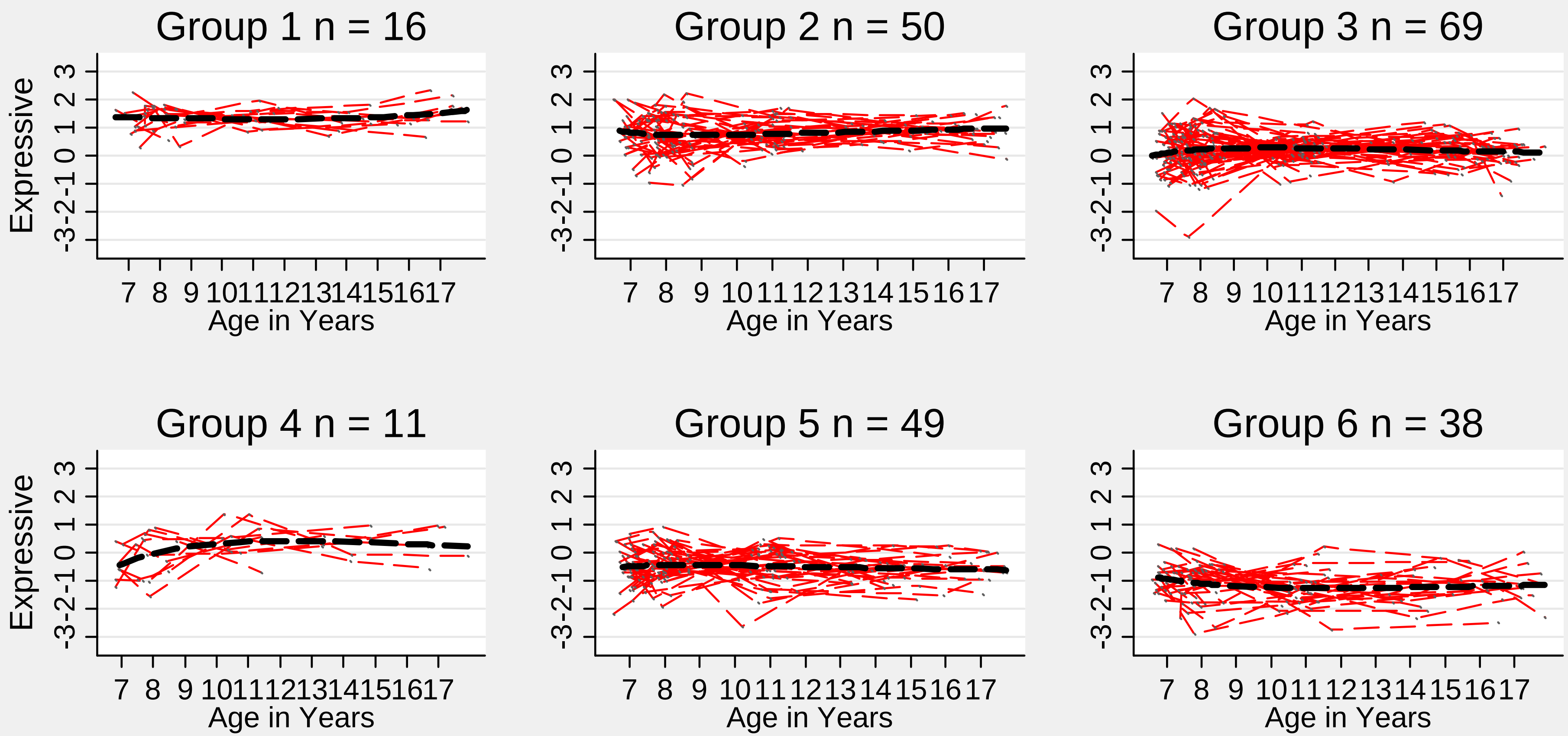
gllamm model for joint expressive and receptive language trajectory classes

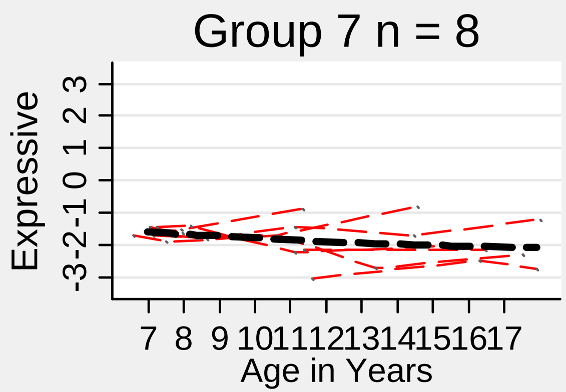
```
eq het: e1 e2 e5 r1 r2 r3 r4  ! Eqn for log std dev of measurement error
eq inte: e1 e2 e5             ! Eqn for expressive intercept factor loadings
eq intr: r1 r2 r3 r4          ! Eqn for receptive intercept factor loadings
eq line: ageye1 ageye2 ageye5        !Eqn for exp linear slope factor loadings
eq linr: ageyr1 ageyr2 ageyr3 ageyr4 !Eqn for rec linear slope factor loadings
cons def 1 [fid1_1l]e2 = [fid1_2l]ageye2 !Constraints for intercept and slope
cons def 2 [fid1_1l]e5 = [fid1_2l]ageye5 !   factor loadings equal
cons def 3 [fid1_3l]r2 = [fid1_4l]ageyr2
cons def 4 [fid1_3l]r3 = [fid1_4l]ageyr3
cons def 5 [fid1_3l]r4 = [fid1_4l]ageyr4

gllamm y e1 e2 e3 e4 r1 r2 r3 r4 ageye1 ageye2 ageyr1, i(fid) nrf(4) /*
 */ eqs(inte line intr linr) s(het) nip(6) cons(1 2 3 4 5) iter(40)/*
 */ nocons trace
```

Classification for categorical latent variables

- Units are usually assigned to latent class with largest posterior probability, often called Maximum Aposteriori (MAP) or Empirical Bayes Modal (EBM)
- Posterior probabilities:

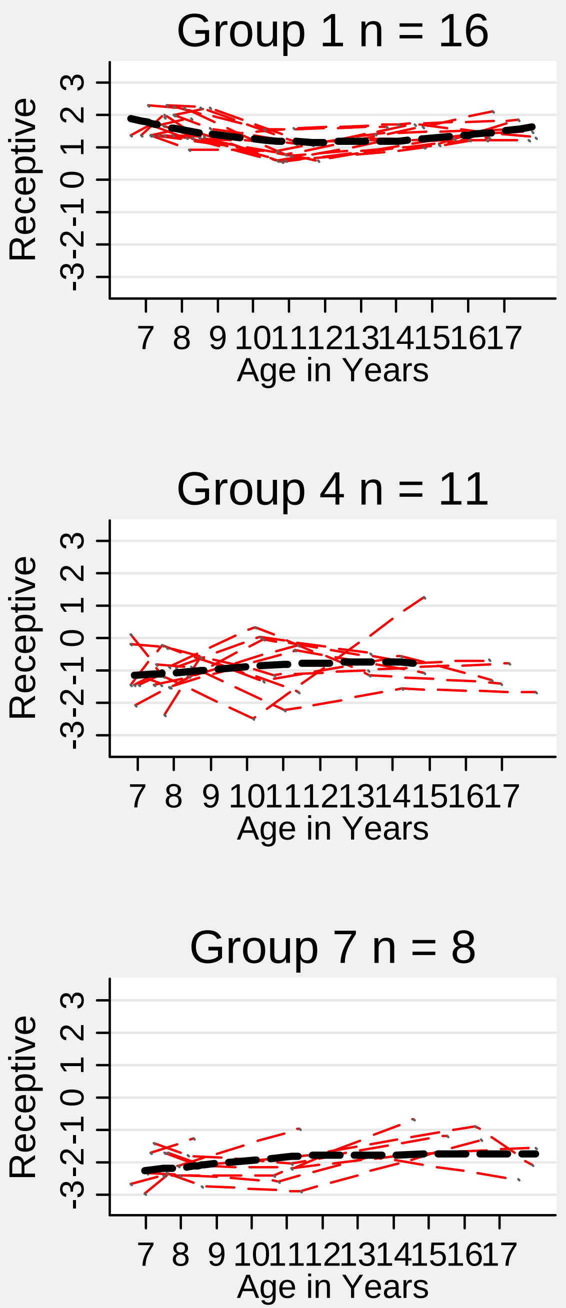
$$\Pr(c \mid y_j) = \frac{\pi_c \prod_{i=1}^{I} \pi_{ij|c}^{y_{ij}} (1 - \pi_{ij|c})^{1-y_{ij}}}{\sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} \pi_{ij|c}^{y_{ij}} (1 - \pi_{ij|c})^{1-y_{ij}}}$$





At each age point (7,8,11,14,16 and 17) the Expressive language score standardized to the mean (0) of the entire SLI population.

No differences in the developmental trajectory, only differences in overall Expressive language ability.



At each age point (7,8,11,14,16 and 17) the Expressive language score standardized to the mean (0) of the entire SLI population.

No differences in the developmental trajectory, only differences in overall Receptive language ability.

End