INCENTIVE EFFECTS IN A NATIONAL PILOT STUDY OF THE OVER FIFTIES IN IRELAND

Mark Hanly, Ian Clifford, Brendan Whelan and George M Savva,
The Irish Longitudinal Study of Ageing
Trinity College Dublin

Non Response Bias

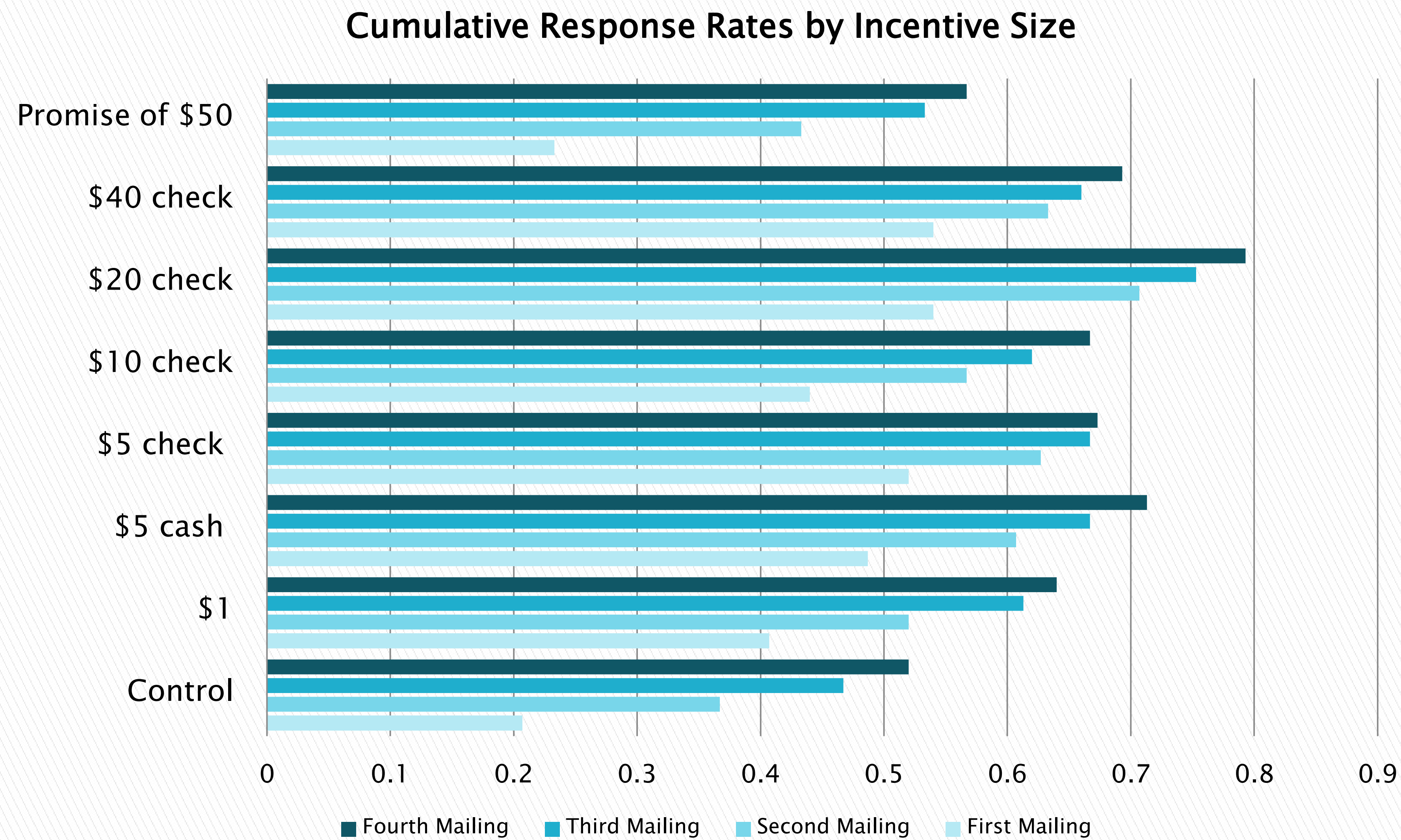
The Bias on an estimate is a function of...

- The Nonresponse Rate (NR)
- The Difference between Respondents and Nonrespondents for that variable

$$Bias\left(\bar{Y}\right) = NR \times \left(Y_{Respondents} - Y_{Nonrespondents}\right)$$

Incentives in social surveys

- Repeatedly shown to increase response
- Cash > gifts/lottery
- Unconditional > Conditional
- Diminishing Returns
- Social Exchange Theory



Data from James and Bolstein (1994)

“Large Monetary Incentives and Their Effect on Mail Survey Response Rates”

But do Incentives decrease bias?

"If response rates are increased using devices that are not **disproportionately attractive to the low propensity groups**, then nonresponse biases may increase despite lowered nonresponse rates"

RM Groves

- Less well off appreciate the money more?
- Rich put more value on their time?

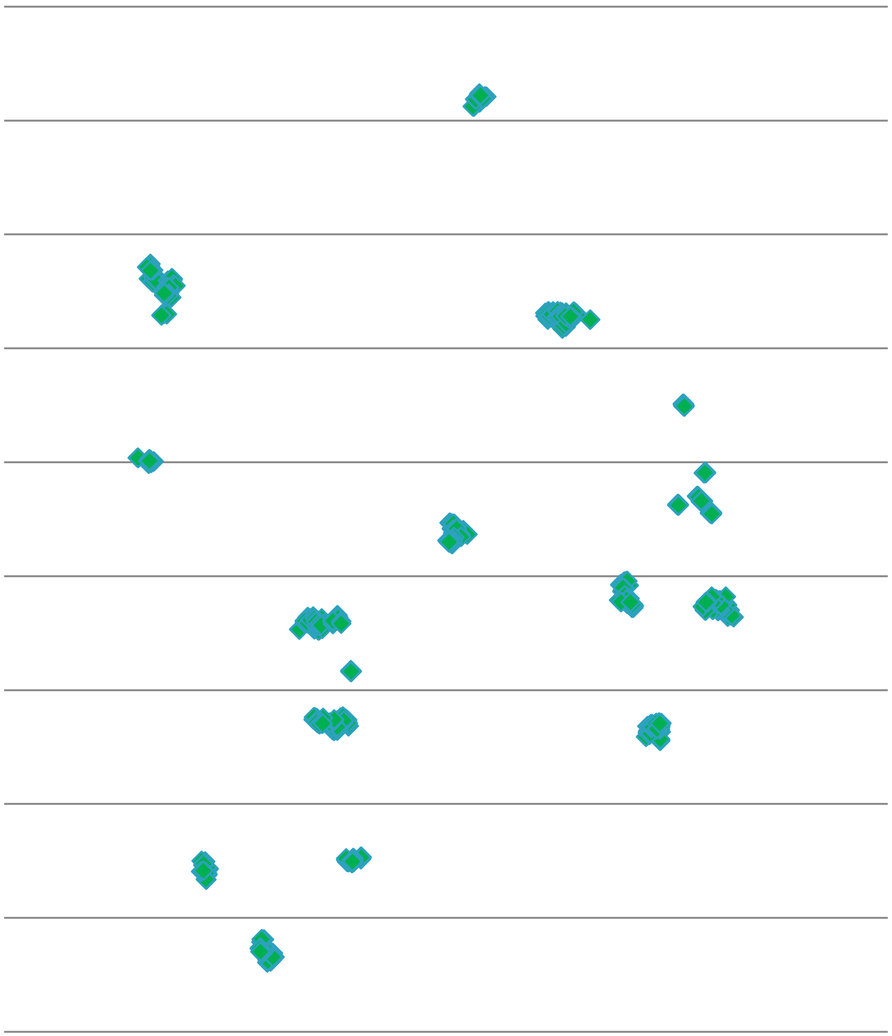
TILDA – The Irish Longitudinal Study of Ageing

- Nationally Representative of Residential population aged 50+
- Pilot study April – June 2009
- Sample Drawn from Geodirectory – Database of all residential addresses

- Three Stage Random Sample
  - 20 geographic clusters
  - 60 Addresses per cluster
  - Next Birthday Rule
- = 1200 households

- Age of householder unknown
- Exact coordinates of household known

TILDA Pilot 2 Sample Addresses



TILDA Wave 1 Sample Addresses

Incentive Experiment

Key Research Questions:

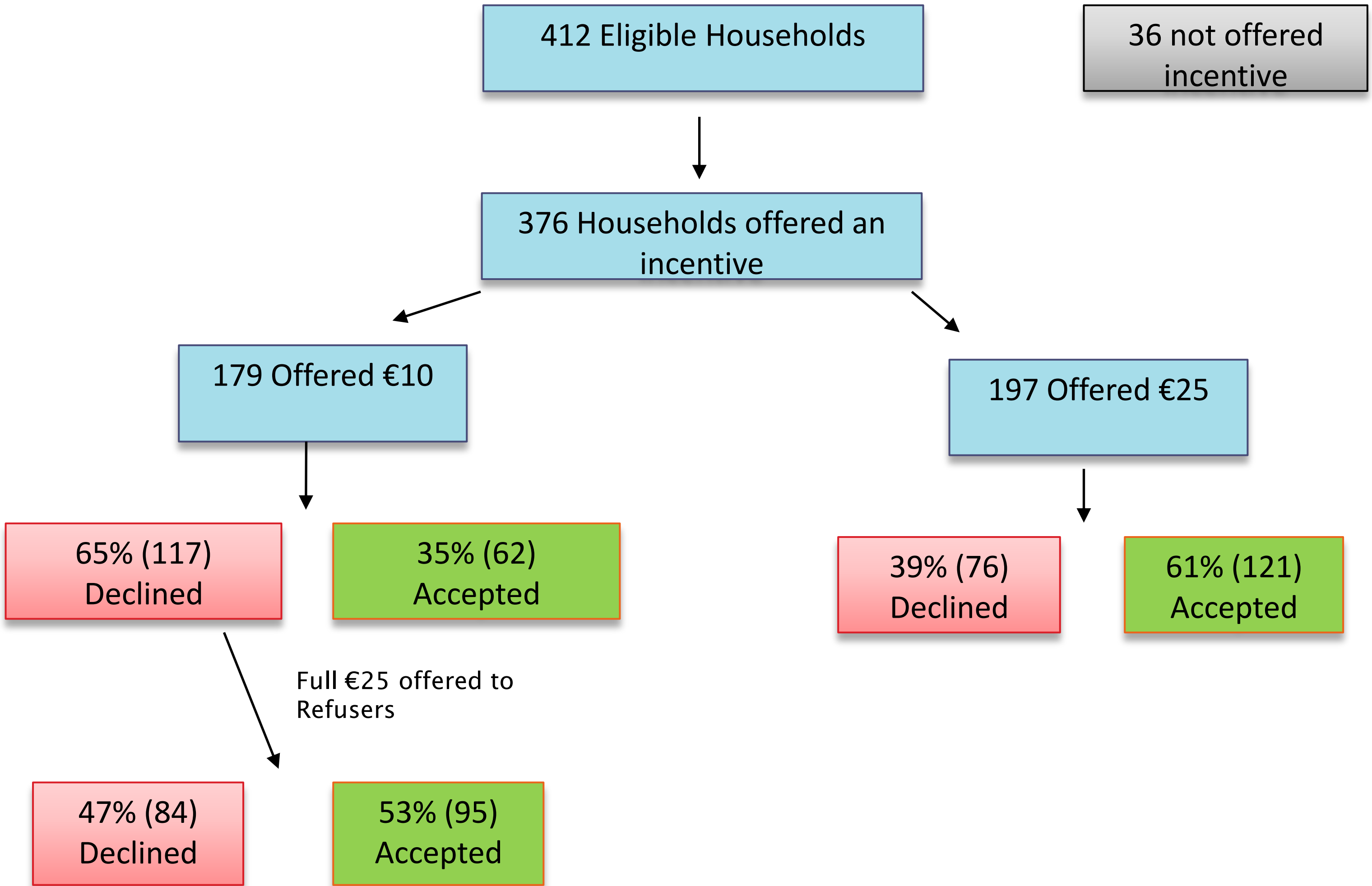
1. Do higher incentives increase response rates?

2. Are incentives more attractive to the less well off?

Incentive Experiment

- Randomises Trial
- Two Conditions: €10 / €25
- 412 Eligible properties identified
- 376 Offered an incentive
  - 179 in €10 Group
  - 197 in €25 Group
- No significant differences between the 2 groups
- Refusers in €10 group offered the full €25

TILDA Pilot Contact Data (N=1200)			
Response Outcome	**n**	**%**	Cuml %
Ineligible property	162	13.5	13.5
Non-contact	90	7.5	21.0
Age ineligible	536	44.7	65.7
Eligible Households	412	34.4	100.0



Flow Diagram for Response to Incentive Experiment

Analysis

- Outcome
  - Response to Initial Offer Binary (0/1)
- Key Predictors
  - Incentive Level Binary (€10/25)
  - Cluster Level Education Standardised Scale
    - % of cluster with 3rd level degree
- Controls

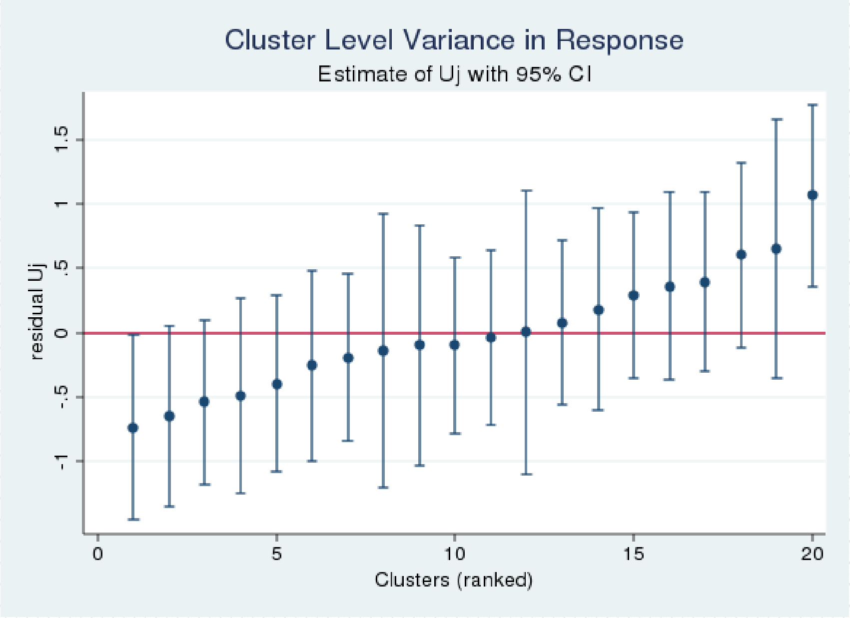
Sex, age, house-type, urbanicity, cluster level occupation profile, cluster level age profile

Modelling 1

- 1. Empty Model With Random Intercept

$$Logit\ (Y_{ij}) = \beta_0 + u_{0j} \qquad u_{0j} \sim N(0, \sigma^2_{u0})$$

	Model 0	Model 1
β_0	-0.532	-0.036
	0.103	*0.181*
$\sigma^2(u_0)$		0.373
		0.209
Deviance	521.0	507.9

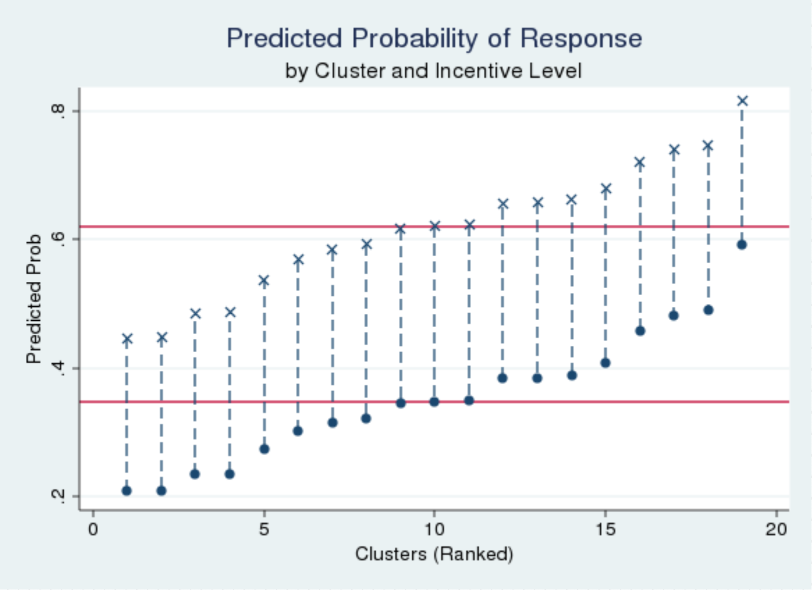


Modelling 2

- 2. Include Incentive Effect

$$Logit\ (Y_{ij}) = \beta_0 + \beta_1(\textbf{Incentive}_{ij}) + u_{0j} \qquad u_{0j} \sim N(0, \sigma^2_{u0})$$

	Model 2
β0	-0.628
	0.217
Incentive (ref €10)	1.121
	0.227
σ2(u0)	0.358
	0.207
Deviance	482.4



Note:
- Dots represent €10 Incentive
- X represents €25 Incentive
- Red lines indicate average response probability for high and low incentive

Modelling 3

- 3. Include All Controls

$$\text{Logit}(Y_{ij}) = \beta 0 + \beta 1\ \text{Incentive}_{ij} + \beta 2\ \text{Sex}_{ij} + \beta 3\ \text{Age}_{ij} + \beta 4\ \text{House-type}_{ij} + \beta 5\ \text{Urbanicity}_{ij} + \beta 6\ \text{Education}_j + \beta 7\ \text{Occupation}_j + \beta 8\ \text{Maturity}_j + u_{0j}$$

$$u_{0j} \sim N(0, \sigma^2_{u0})$$

	Model 1	Model 2	Model 3
β_0	-0.036	-0.628	0.011
Incentive (ref €10)		1.121***	1.351***
Sex (ref male)			0.300
Age (ref 50s)			
60s			-0.126
70s			-0.114
House Type (ref detached)			-0.173
Urbanicity (ref Dublin)			
Other Urban			-0.884
Rural			-0.626
Education			-0.245
Occupation			0.259
Maturity			0.053
$\sigma^2(u_0)$	0.373	0.358	0.294
Deviance	507.9	482.4	438.8

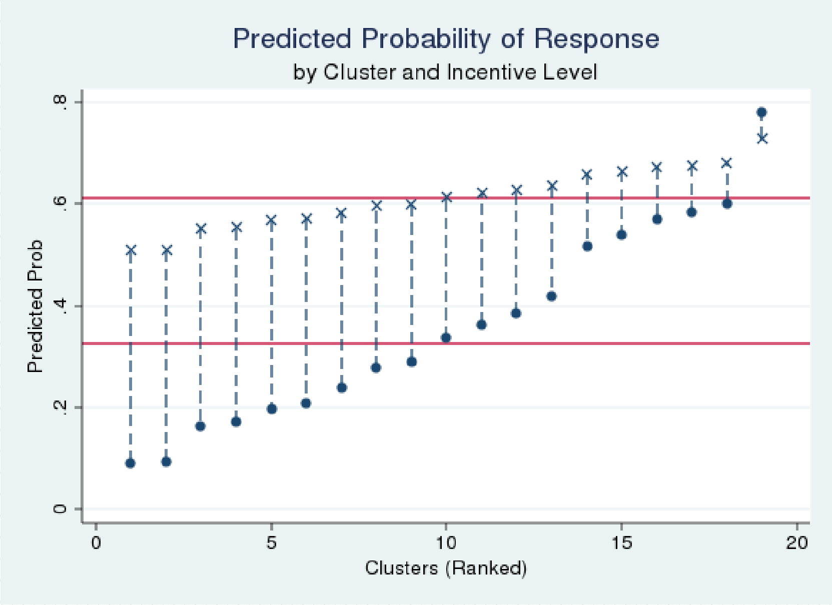
Modelling 4

- 4. Allow for random effect of incentive

Logit (Y_{ij}) $=\beta 0 + \beta 1$ ***Incentive***$_{ij}$ + u0j + u_{1j} ***Incentive***$_{ij}$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}\right\}$$

	Model 4
Intercept	-0.720
	0.338
Incentive	1.170
	0.317
σ2(u0)	1.402
	0.821
σ2(u1)	0.755
	0.619
σu01	-1.030
	0.70
Deviance	474.3



Note:
– Dots represent €10 Incentive
– X represents €25 Incentive
– Red lines indicate average response probability for high and low incentive

Modelling 5

- 5. Can a clusters education profile explain the random effect of incentive?

$$\text{Logit }(Y_{ij}) = \beta_0 + \beta_1 \textbf{\textit{Incentive}}_{ij} + \beta_2 \textbf{\textit{Education}}_j + \beta_3 \textbf{\textit{Incentive}}_{ij} * \textbf{\textit{Education}}_j + u_{0j} + u_{1j} \textbf{\textit{Incentive}}_{ij}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}\right\}$$

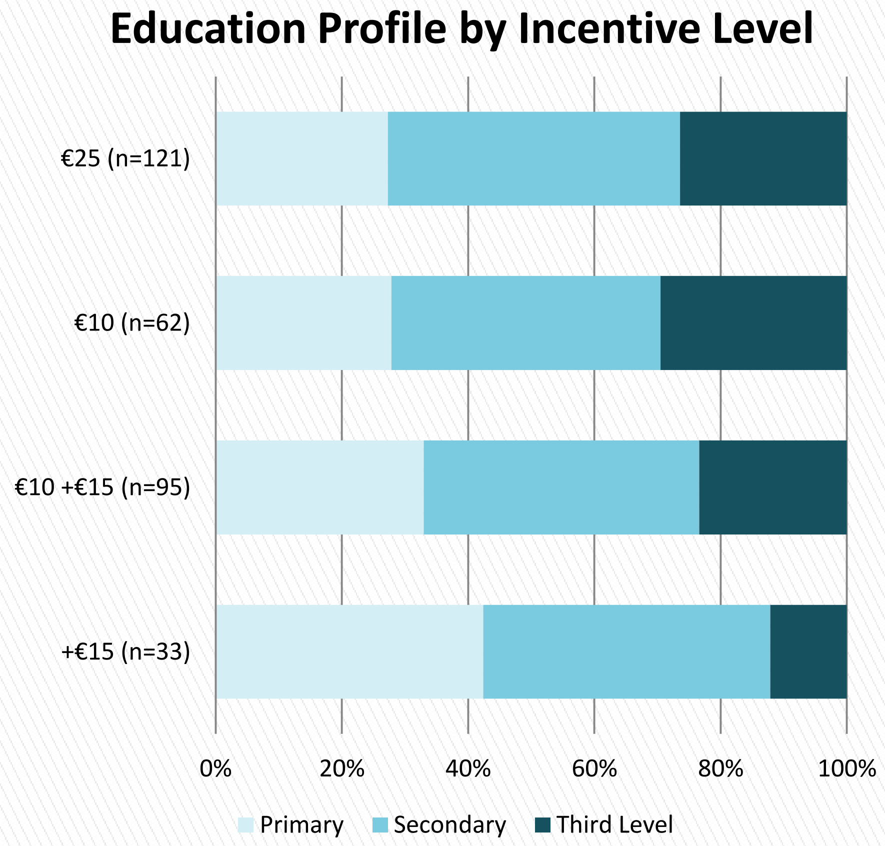
	Model 4	Model 5
Intercept	-0.720	-0.763
Incentive	1.170***	1.210***
Education		0.363
Education*Incentive		-0.186
$\sigma^2(u_0)$	1.402	1.330
$\sigma^2(u_1)$	0.755	0.713
$\sigma\ (u_{01})$	-1.030	-0.974
Deviance	474.3	472.4

Conclusions

- Clear significant effect of the higher incentive amount
- Significant variation across clusters but variation could not be explained in term of cluster education, occupation or maturity profile
- No evidence that incentives work better in poorer clusters

Epilogue

- But...
- Recall that refusers at €10 were offered the full €25 to participate
- 33 individuals were persuaded by this offer
- Some evidence that these individuals had lower educational status



Thank you for your Attention

Heritability and environmentality of intelligence subscales

Genetic model

We can exploit the different genetic relatedness of identical (MZ) and non-identical (DZ) twins to decompose variation in an outcome into a genetic, a shared environmental, and a non-shared environmental component).

Model

$$y_{ij} = \beta_0 + \beta_1 x_j + c_j + a_{ij} + e_{ij}$$

$$\begin{aligned} c_j &\sim \mathrm{N}(0, \sigma^2_C) \\ a_{ij} &\sim \mathrm{N}(0, \sigma^2_A) \\ e_{ij} &\sim \mathrm{N}(0, \sigma^2_E) \end{aligned}$$

$$\mathrm{cov}(g_{1j}, g_{2j}) = r_j \sigma^2_G$$

$$r_j = \begin{cases} 1 & \text{MZ twins} \\ 0.5 & \text{DZ twins} \end{cases}$$

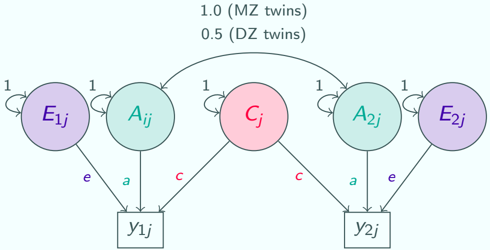
Fitting the model - option 1

$$\textbf{score}_{1j} = \beta_0 + \beta_1 \textbf{score}_{2j} + \beta_2 \textbf{r}_j + \beta_3 \textbf{r.score}_{2j} + e_j$$

DeFries-Fulker regression

- Fits a single level regression
- One twin's outcome is the response and the other twin's outcome an explanatory variable in various interactions
- Proportions of genetic and shared environmental variance given by β_3 and β_1 respectively
- Useful for restricting to range of outcome values of interest but not very extensible.

Fitting the model - option 2



Structural equation modelling

- Genetic, shared environmental, and non-shared environmental factors are latent variables with variances constrained to 1
- Components given by squares of respective loadings
- Twins' outcomes in separate variables
- Correlation between genetic latent variables constrained to be 1 for MZ and 0.5 for DZ twins

Fitting the model - option 3

Multilevel genetic model

- Two level model but with 3 (sets of) random effects
- Need reparameterisation to get software to fit it
- Rabe-Hesketh et al. (2008) provide two
  - fit a two level model with genetic effect for DZ twins split into shared (level 2) and non-shared (level 1) part, with variances constrained to be equal
  - or a three level model (level 1 = individual, level 3 = family, level 2 = individual (DZ) or family (MZ)) then desired components given by linear combinations of estimated variances

We use a multilevel genetic model with the first reparameterisation

Fitting the model - model used

First reparameterisation

$$y_{ij} = \beta_0 + u_j + g_{1j}\left(\mathbf{MZ}_j + \sqrt{0.5}\mathbf{DZ}_j\right) + g_{2ij}\left(\sqrt{0.5}\mathbf{DZ}_j\right) + e_{ij}$$

$$u_j \sim \mathrm{N}(0, \sigma^2_C)$$
$$g_{1j} \sim \mathrm{N}(0, \sigma^2_A)$$
$$g_{2ij} \sim \mathrm{N}(0, \sigma^2_A)$$
$$e_{ij} \sim \mathrm{N}(0, \sigma^2_E)$$

- Started with this model
- Likelihood ratio tests used to assess significance of genetic and shared environmental component
- Dropped these from model if not significant
- Results presented are for final, best model.

Data

National Collaborative Perinatal Project

- 12 different urban sites in the US participated
- Pregnancies between 1959 and 1966
- At each site, all pregnant women or a random selection of them sampled
- For sampled women, subsequent pregnancies within study timeframe also included
- c. 50,000 pregnancies of which 504 twin pairs with zygosity information (1008 twins)
- We look at subscales of the Wechsler Intelligence Scales for Children (WISC), administered at age 7
- We also have information on SES at birth and at age 7 (education, occupation, income)
- 575 of the 1008 have all IQ subscales and all SES measures; 6 have none of this data

Measures: intelligence subscales

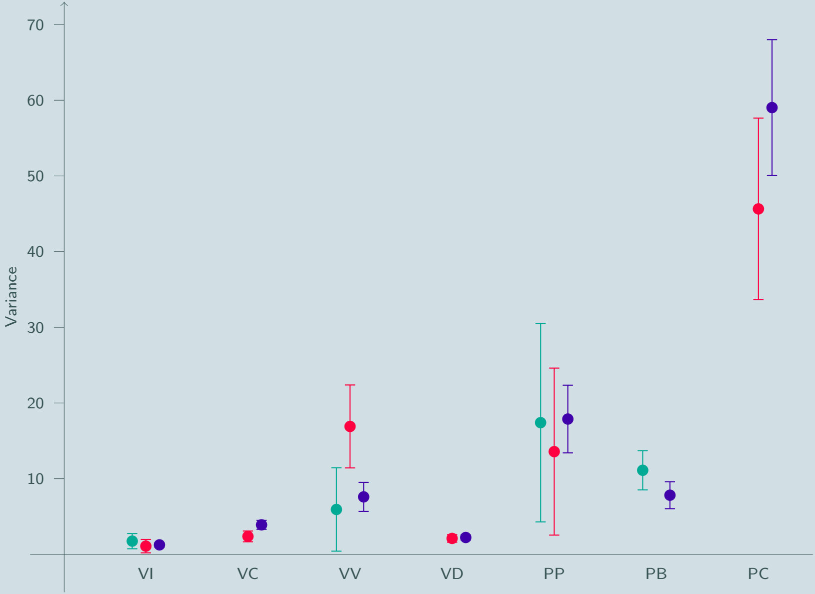
	Verbal	
Information	VI	General knowledge, e.g. "How many days are there in a week?". 30 items.
Comprehension	VC	Knowledge and judgement in social and practical situations, e.g. "What would you do if you cut your finger?". 14 items.
Vocabulary	VV	Definitions of words, e.g. "What is a bicycle?". 35 items.
Digit Span	VD	Working memory. Repeat a sequence of digits read out by the examiner, in the same order (14 items, from 3 to 9 digits long) or in reverse order (14 items, from 2 to 8 digits long).
	Performance	
Picture Arrangement	PP	Put picture cards in correct order to form a story. 7 items(?)
Block Design	PB	Use red and white blocks to recreate a presented design. 7 items(?)
Coding	PC	Add particular marks to particular shapes: one vertical line inside every star, two horizontal lines inside every circle, one horizontal line inside every triangle, a circle inside every cross, and two vertical lines inside every square. 45 items.

Measures: socioeconomic status variables

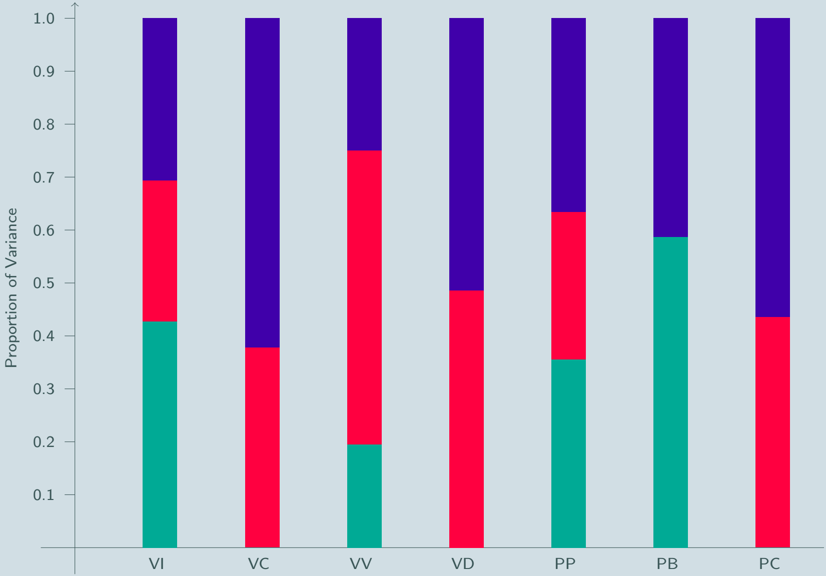
At birth		
Education (internal coding)	e01	0-9
Occupation (internal coding)	o01	0-9
Income (internal coding)	i01	Binary
Education (US Census Bureau coding)	e02	1-9
Occupation (US Census Bureau coding)	o02	1-9
Income (US Census Bureau coding)	i02	Binary
Age 7		
Education	e7	0-97
Occupation	o7	0-94
Income	i7	1-98

Which parent used varies from family to family. Income adjusted to be comparable across all enrollment dates. About 8% of the twins have $\mathbf{i01} = 1$ and about 3% have $\mathbf{i02} = 1$

Results: with no explanatory variables



As proportions



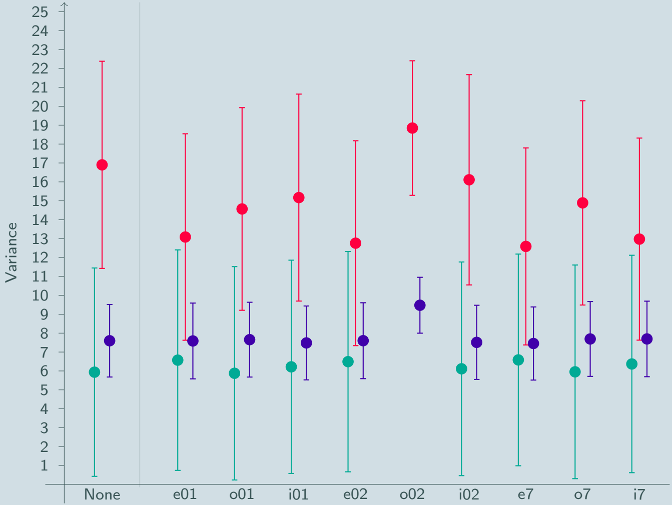
Adding explanatory variables

- Can some of the socioeconomic variables explain some of the shared environmental variation?
- Can they ‘explain’ some of the genetic variation (through genetic correlation)?
- Again, used likelihood ratio tests to identify best model, discarding shared environmental and/or genetic variation where non-significant

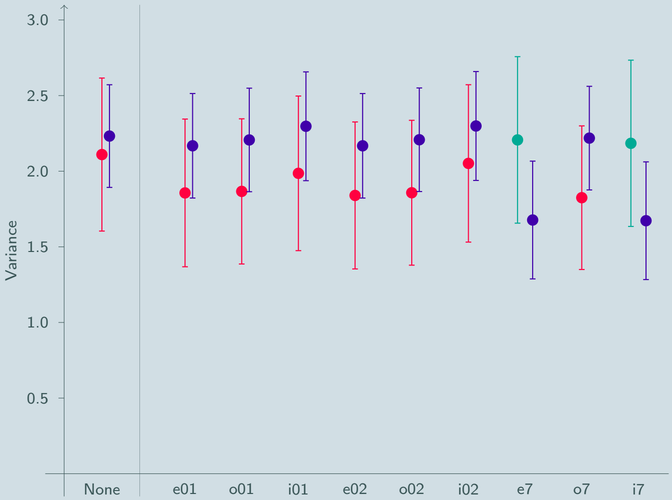
Verbal Information

Verbal Comprehension

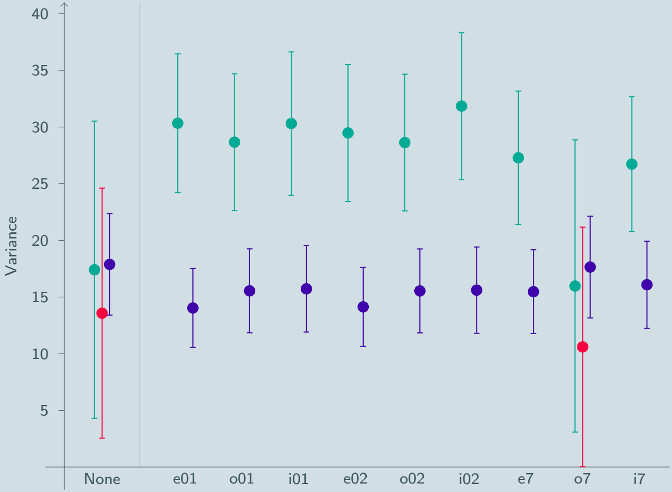
Verbal Vocabulary



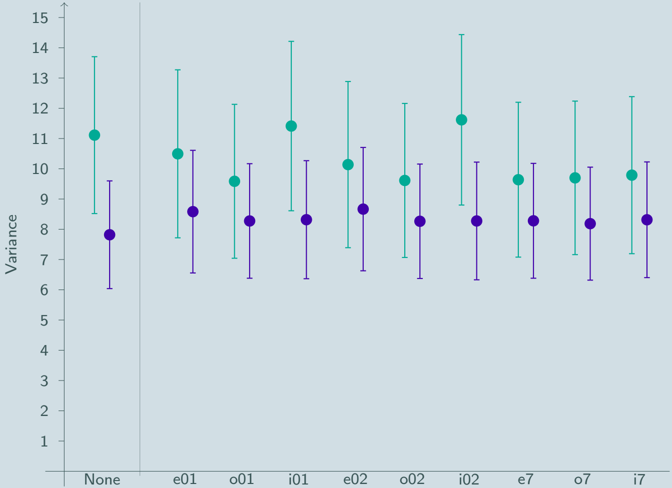
Verbal Digit Span



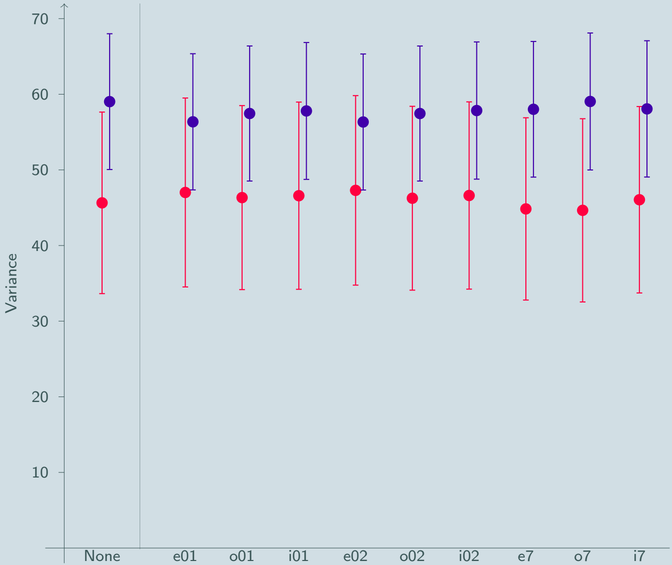
Performance Picture Arrangement



Performance Blocks



Performance Coding



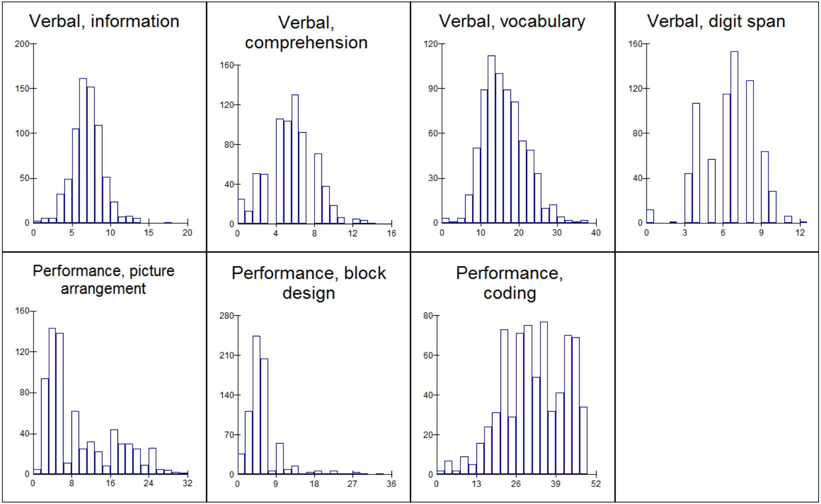
Future work

- Deal with missing data (impute)
- Check distributional assumptions
- Look at differential heritability and environmentality
- See how much of each component each subscale contributes to the full IQ score

Extras

- Distributions of variables
- Missing data patterns
- Second reparameterisation of multilevel model
- Parameter estimates for no-explanatory-variable models
- Model for differential heritability and environmentality

Distribution of the 7 scales



Some of these are clearly not very Normal.

Verbal info	Verbal comp	Verbal vocab	Verbal digit	Perf pic	Perf block	Perf coding	Education01	Occupation01	Income01	Education02	Occupation02	Income02	Education7	Occupation7	Income7	Number
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	6
■	■	■	■	■	■	■	□	□	□	□	□	□	■	■	■	186
■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	58
■	■	■	■	■	■	■	□	□	■	□	□	■	■	■	■	16
■	■	■	■	■	■	■	□	■	□	□	■	□	■	■	■	4
□	□	□	□	□	□	□	■	□	□	■	□	□	□	□	□	44
□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	22
□	□	□	□	□	□	□	■	■	■	■	■	■	□	□	□	10
□	□	□	□	□	□	□	□	■	□	□	■	□	□	□	□	8
□	□	□	□	□	□	□	□	■	□	□	■	□	■	■	■	4
□	□	□	□	□	□	□	■	□	■	■	□	■	□	□	□	2
□	□	□	□	□	□	□	■	■	■	■	■	■	■	□	□	2
□	□	□	□	□	□	□	■	□	■	■	□	■	□	□	■	2
□	□	□	□	□	□	□	■	■	□	■	■	□	□	□	□	2
□	□	□	□	□	□	□	□	□	□	□	□	□	■	□	□	2
■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	■	1
□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	11
■	■	■	■	■	■	■	■	□	□	■	□	□	□	□	■	1
□	□	□	□	□	□	□	■	□	□	■	□	□	□	□	■	1
■	■	■	■	■	■	■	□	□	■	□	□	■	□	□	□	5
□	□	□	□	□	□	□	□	□	■	□	□	■	□	□	□	27
■	■	■	■	■	■	■	■	□	□	■	□	□	■	■	■	9
□	□	□	□	□	□	□	■	□	□	■	□	□	■	■	■	1
■	■	■	■	■	■	■	□	■	□	□	■	□	□	□	■	1
□	□	□	□	□	□	□	□	■	□	□	■	□	□	□	■	1
■	■	■	■	■	■	■	■	□	□	■	□	□	□	□	□	5
□	□	■	□	□	□	□	■	□	□	■	□	□	□	□	□	1
□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□	1
□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	575

Second reparameterisation

Model

$$y_{ijk} = \beta_0 + u_k + g_{jk} + e_{ijk}$$

$$\begin{aligned} u_k &\sim \mathrm{N}(0, \sigma_u^2) \\ g_{jk} &\sim \mathrm{N}(0, \sigma_g^2) \\ e_{ijk} &\sim \mathrm{N}(0, \sigma_e^2) \end{aligned}$$

$$\begin{aligned} \sigma_C^2 &= \sigma_u^2 - \sigma_g^2 \\ \sigma_A^2 &= 2\sigma_g^2 \\ \sigma_E^2 &= \sigma_e^2 \end{aligned}$$

Basic model

$\text{verbal\_info}_{ij} \sim \mathrm{N}(XB,\ \Omega)$

$\text{verbal\_info}_{ij} = \beta_{0ij}\text{cons} + e_{2ij}\text{DZ}_j + u_{1j}\text{MZDZ}_j$

$\beta_{0ij} = 6.554(0.092) + u_{0j} + e_{0ij}$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 1.091(0.451) & \\ 0 & 1.752(0.513) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \\ e_{2ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 1.258(0.161) & \\ 0 & 1.752(0.513) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 2909.273(716 of 1008 cases in use)

$\text{verbal\_comp}_{ij} \sim \mathrm{N}(XB,\ \Omega)$

$\text{verbal\_comp}_{ij} = \beta_{0ij}\text{cons}$

$\beta_{0ij} = 5.501(0.108) + u_{0j} + e_{0ij}$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 2.377(0.362) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 3.903(0.302) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3296.479(716 of 1008 cases in use)

$\text{verbal\_vocab}_{ij} \sim \mathrm{N}(XB,\ \Omega)$

$\text{verbal\_vocab}_{ij} = \beta_{0ij}\text{cons} + e_{2ij}\text{DZ}_j + u_{1j}\text{MZDZ}_j$

$\beta_{0ij} = 15.723(0.261) + u_{0j} + e_{0ij}$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 16.901(2.795) & \\ 0 & 5.935(2.812) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \\ e_{2ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 7.597(0.978) & \\ 0 & 5.935(2.812) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 4256.022(715 of 1008 cases in use)

$\text{verbal\_digit}_{ij} \sim \mathrm{N}(XB,\ \Omega)$

$\text{verbal\_digit}_{ij} = \beta_{0ij}\text{cons}$

$\beta_{0ij} = 6.381(0.093) + u_{0j} + e_{0ij}$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 2.110(0.258) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 2.232(0.173) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 2990.387(715 of 1008 cases in use)

$\text{perf\_pic}_{ij} \sim \mathrm{N}(XB,\ \Omega)$

$\text{perf\_pic}_{ij} = \beta_{0ij}\text{cons} + e_{2ij}\text{DZ}_j + u_{1j}\text{MZDZ}_j$

$\beta_{0ij} = 9.499(0.315) + u_{0j} + e_{0ij}$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 13.577(5.629) & \\ 0 & 17.406(6.692) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \\ e_{2ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 17.880(2.282) & \\ 0 & 17.406(6.692) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 4706.085(716 of 1008 cases in use)

$\text{perf\_block}_{ij} \sim \mathrm{N}(XB,\ \Omega)$

$\text{perf\_block}_{ij} = \beta_{0i}\text{cons} + e_{2ij}\text{DZ}_j + u_{1j}\text{MZDZ}_j$

$\beta_{0i} = 5.934(0.189) + e_{0ij}$

$$\begin{bmatrix} u_{1j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 11.113(1.323) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \\ e_{2ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 7.819(0.908) & \\ 0 & 11.113(1.323) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 4068.284(716 of 1008 cases in use)

$\text{perf\_coding}_{ij} \sim \mathrm{N}(XB,\ \Omega)$

$\text{perf\_coding}_{ij} = \beta_{0ij}\text{cons}$

$\beta_{0ij} = 31.768(0.451) + u_{0j} + e_{0ij}$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 45.643(6.124) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 59.027(4.580) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 5292.296(716 of 1008 cases in use)

Differential heritability and environmentality

$$y_{ij} = \beta_0 + \beta_1 x_j + c_{0j} + c_{1j} x_j + a_{0ij} + a_{1ij} x_j + e_{0ij} + e_{1ij} x_j$$

$$\begin{bmatrix} c_{0j} \\ c_{1j} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{C0} & \\ \sigma_{C01} & \sigma^2_{C1} \end{bmatrix} \right)$$

$$\begin{bmatrix} a_{0ij} \\ a_{1ij} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{A0} & \\ \sigma_{A01} & \sigma^2_{A1} \end{bmatrix} \right)$$

$$\begin{bmatrix} e_{0ij} \\ e_{1ij} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{E0} & \\ \sigma_{E01} & \sigma^2_{E1} \end{bmatrix} \right)$$

$$\mathrm{cov}(a_{01j}, a_{02j}) = r_j \sigma^2_{A0}$$

$$\mathrm{cov}(a_{11j}, a_{12j}) = r_j \sigma^2_{A1}$$

$$\mathrm{cov}(a_{01j}, a_{12j}) = r_j \sigma_{A01}$$

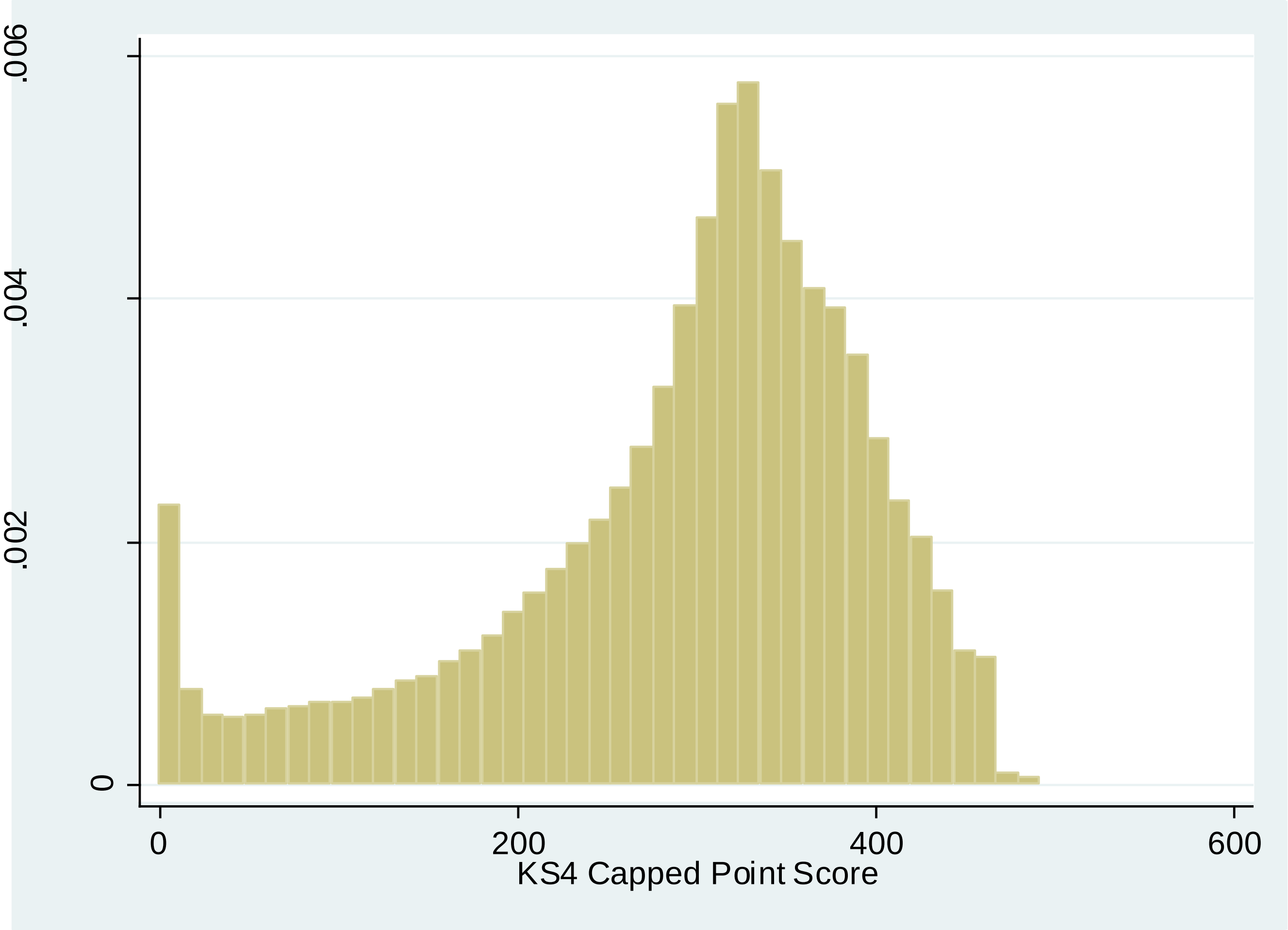
$$\mathrm{cov}(a_{11j}, a_{02j}) = r_j \sigma_{A01}$$

USING LINKED CENSUS AND SURVEY DATA TO IDENTIFY BIAS IN MEASURES OF CHILD'S HOME BACKGROUND AND THE IMPLICATIONS FOR SCHOOL EVALUATION

Rob French

RSS Avon Group 28th March 2012

Outcome: KS4 Capped Point Score



School evaluation model:

$$y_{ij} = \beta_1 + \beta_2 \chi_{2ij} + \cdots + \beta_p \chi_{pij} + \beta_q FSM_{qij} + \zeta_j + \epsilon_{ij}$$

Standard controls:

- prior attainment,
- Special Educational Needs (SEN),
- Income Deprivation Affecting Children Index (IDACI),
- Free School Meals (FSM),
- student mobility,
- gender, in care,
- ethnicity,
- English as an Additional Language (EAL),
- Age.

(FSM ethnicity interactions are omitted)

School residuals

Individual residuals

Variation in KS4CPS explained

Single Variable Model	R Squared
Prior Attainment (KS2)	0.463
Special Educational Needs (SEN)	0.266
Local Area Deprivation (IDACI)	0.092
Free School Meals (FSM)	0.062
Pupil Mobility: Started Outside Jul-Sep Yr. 7-9	0.041
Pupil Mobility: Started After Yr. 10	0.021
Female	0.017
In Care	0.010
Ethnicity	0.010
English Additional Language (EAL)	0.001
Age (DOB)	0.001

Value Added Estimates for the Entire Cohort in 2006 (Population)

	beta	s.e.		beta	s.e.
KS2: Average Point Score	12.30***	-0.0293	Ethnicity: Any Other Mixed Background	5.123***	-1.118
KS2: Average Point Score Squared	0.359***	-0.00408	Ethnicity: Any Other White Background	11.66***	-0.815
KS2: Maths Deviation from KS2APS	1.612***	-0.0517	Ethnicity: Bangladeshi	30.73***	-1.304
KS2: English Deviation from KS2APS	2.077***	-0.0507	Ethnicity: Black African	30.68***	-1.05
Female	14.00***	-0.207	Ethnicity: Caribbean	12.84***	-0.882
Eligible for Free School Meals	-21.08***	-0.313	Ethnicity: Chinese	32.36***	-1.832
English Additional Language	18.16***	-0.68	Ethnicity: Gypsy Romany	-49.96***	-4.472
SEN: None	ref.		Ethnicity: Indian	22.34***	-0.903
SEN: Action	-45.95***	-0.352	Ethnicity: Irish	-2.311	-1.608
SEN: Action Plus/Statement	-71.79***	-0.418	Ethnicity: Pakistani	20.65***	-0.953
In Care	-31.14***	-1.183	Ethnicity: Traveller of Irish Heritage	-61.58***	-6.316
Index of Deprivation Affecting Children	-68.28***	-0.734	Ethnicity: White and Asian	10.71***	-1.51
Date of Birth (Centred)	-11.05***	-0.331	Ethnicity: White and Black African	9.535***	-2.223
Mobility: After Year 10 Start	-48.34***	-0.801	Ethnicity: White and Black Caribbean	-1.172	-1.035
Mobility: Outside Jul-Sep, Yr 7-9	-25.67***	-0.417	School KS2 Mean	6.412***	-0.133
Ethnicity: White British	ref.		School KS2 Standard Deviation	-1.115*	-0.441
Ethnicity: Any Other Asian Background	26.36***	-1.437	Constant	213.1***	-2.98
Ethnicity: Any Other Black Background	10.37***	-1.533	Between School SD	29.91***	-0.422
Ethnicity: Any Other Ethnic Group	27.69***	-1.32	Within School SD	67.97***	-0.0671

	beta	s.e.
KS2: Average Point Score	12.30***	-0.0293
KS2: Average Point Score Squared	0.359***	-0.00408
KS2: Maths Deviation from KS2APS	1.612***	-0.0517
KS2: English Deviation from KS2APS	2.077***	-0.0507
English Additional Language	18.16***	-0.68
Date of Birth (Centred)	-11.05***	-0.331

	beta	s.e.
School KS2 Mean	6.412***	-0.133
School KS2 Standard Deviation	-1.115*	-0.441
Between School SD (sigma u)	29.91***	-0.422
Within School SD (sigma e)	67.97***	-0.0671

Free School Meals (FSM)

- income support
- income based job-seekers allowance
- child tax credit (but not WTC & income <14k)
- support under part IV of the immigration and asylum act 1999
- the guarantee element of state pension credit

Data

The 2 linked datasets:

- National Pupil Database (NPD) a census
- Longitudinal Study of Young People in England (LSYPE) a survey

Analysis: 3 models

- **Omitted variable model**, CVA with no home background measure

$$y_{ij} = \beta_1 + \beta_2 \chi_{2ij} + \cdots + \beta_p \chi_{pij} + \zeta_j + \epsilon_{ij}$$

- **Better measures model**, CVA with better measures of home background (from survey data)

$$y_{ij} = \beta_1 + \beta_2 \chi_{2ij} + \cdots + \beta_p \chi_{pij} + \beta_r \mathrm{Income}_{rij} + \zeta_j + \epsilon_{ij}$$

- **Proxy model**, CVA model with FSM as a proxy for home background (from census data)

$$y_{ij} = \beta_1 + \beta_2 \chi_{2ij} + \cdots + \beta_p \chi_{pij} + \beta_q FSM_{qij} + \zeta_j + \epsilon_{ij}$$

Identifying the **total** Omitted Variable Bias from not including a measure of household deprivation, using the **best available** measures of home background

	Omitted Variable Model [1]		Better Measures Model [2]		Difference [1-2]	**TOTAL** Omitted Variable Bias [(1-2)/1]
FSM	N		N			
IDACI	-62.79***	(3.994)	-36.36***	(4.067)	-26.43	42%
CVA	Y		Y			
Income	N		**Y**			
NSSEC	N		**Y**			
Education	N		**Y**			
Constant	206.1***	(16.36)	161.4***	(18.59)		
Sigma u	18.14***	(0.977)	17.32***	(0.954)	0.830	4.6%
Sigma e	66.67***	(0.427)	65.01***	(0.417)		
Rho	0.0689		0.0663			
R Squared	0.592		0.613		* 0.0212	3.6%
Observations	13133		13133			

*Absolute difference in R squared

Identifying the **amount** of Omitted Variable Bias from not including a measure of household deprivation, using a **proxy** for home background

	Omitted Variable Model [1]		Proxy Model [3]		Difference [1-3]	Omitted Variable Bias identified by **PROXY** [(1-3)/3]
FSM	N		-20.21***	(1.752)		
IDACI	-62.79***	(3.994)	-53.31***	(4.058)	-9.48	15.1%
CVA	Y		Y			
Income	N		**N**			
NSSEC	N		**N**			
Education	N		**N**			
Constant	206.1***	(16.36)	209.2***	(16.28)		
Sigma u	18.14***	(0.977)	18.08***	(0.971)	0.060	0.3%
Sigma e	66.67***	(0.427)	66.33***	(0.425)		
Rho	0.0689		0.0692			
R Squared	0.592		0.596		* 0.0041	0.7%
Observations	13133		13133			

*Absolute difference in R squared

Finding the proportion of the **total** bias that is eliminated by the **proxy**

	Omitted Variable Model [1]	Better Measures Model [2]	Proxy Model [3]	**TOTAL** Omitted Variable Bias [(1-2)/1]	Omitted Variable Bias identified by **PROXY** [(1-3)/3]
FSM	N	**N**	-20.21***		
IDACI	-62.79***	-36.36***	-53.31***	42%	15.1%
CVA	Y	Y	Y		
Income	N	**Y**	**N**		
NSSEC	N	**Y**	**N**		
Education	N	**Y**	**N**		
Constant	206.1***	161.4***	209.2***		
Sigma u	18.14***	17.32***	18.08***	4.6%	0.3%
Sigma e	66.67***	65.01***	66.33***		
Rho	0.0689	0.0663	0.0692		
R Squared	0.592	0.613	0.596	3.6%	0.7%
Observations	13133	13133	13133		

Conclusions

FSM a poor proxy in the school evaluation context

- It does not explain a substantial proportion of the omitted variable bias (from omitting home background measures) for accurate estimation of other covariates
- It does not explain a substantial proportion of the variance between schools resulting from differences in the home background of a schools intake
- It does not explain a substantial proportion of the variation attributable to home background (when better measures are used)

Literature

School Evaluation:

- Ben Stanbury (2008) Contextualised Value Assed Models in the English Educational System

Papers that try to evaluate FSM as a proxy:

- Daphne Kounali, Tony Robinson, Harvey Goldstein and Hugh Lauder (2007) The probity of free school meals as a proxy measure for disadvantage
- Graham Hobbs & Anna Vignoles (2007) Is Free School Meal Status a valid Proxy for Socio-Economic Status in Schools Research

Summary of Models Used for Each of the Three Samples

	NPD Population Sample	LSYPE Schools Sample	LSYPE Participants Sample	
Descriptive Statistics	Y	Y	Y	
Null OLS Model	Y	Y	Y	$y_{ij} = \beta + \xi_{ij}$
Individual Context OLS Model	Y	Y	Y	$y_{ij} = \beta_1 + \beta_2\chi_{2ij} + \cdots + \beta_p\chi_{pij} + \xi_{ij}$
Null RE Model	Y	Y	Y	$y_{ij} = \beta_1 + \zeta_j + \epsilon_{ij}$
Individual Context RE Model	Y	Y	Y	$y_{ij} = \beta_1 + \beta_2\chi_{2ij} + \cdots + \beta_p\chi_{pij} + \zeta_j + \epsilon_{ij}$
Individual & School Context RE Model	Y	Y	Y	$y_{ij} = \beta_1 + \beta_2\chi_{2ij} + \cdots + \beta_p\chi_{pij} + \zeta_j + \epsilon_{ij}$

Null OLS Regressions

	NPD Sample		LSYPE Schools Sample		LSYPE Participants Sample (Weighted)		LSYPE Participants Sample (Unweighted)	
Constant	290.6***	(0.150)	287.9***	(0.268)	292.5***	(1.066)	296.2***	(0.906)
Observations	520296		157718		10313		13133	

Null Random Effects Regressions for all Three Samples

	NPD Sample		LSYPE Schools Sample		LSYPE Participants Sample	
Constant	244.6***	(1.530)	270.4***	(2.115)	280.6***	(2.240)
Between School SD	107.8***	(1.165)	67.89***	(1.616)	57.27***	(2.101)
Within School SD	93.28***	(0.0920)	96.33***	(0.172)	91.75***	(0.596)
Rho	0.572		0.332		0.280	
Observations	520296		157718		13133	
Standard errors in parentheses, * p < 0.05, ** p < 0.01, *** p < 0.001						