Synthetic Cohort Construction using linked Administrative data with the UKHLS

Dr Scott Oatley

Research Associate

English Longitudinal Study of Ageing (ELSA)

University of Manchester

Focus

- Forms a wider research agenda
  - Study of youth transitions
- Really two presentations in one
  - The first focuses on the construction of synthetic birth cohorts using non-birth cohort social survey data
  - The second relates to the complications of working with linked administrative data in a social survey setting
- Presentation will finish with a more sociologically substantive point (You get to see some nice graphs)

Background

- Young people's transition to adulthood
  - First major transition from school-to-work

- The study of a young person's first transition at age 16
  - Staying within education or not

- **How do the social processes and social inequalities that influence young people's transitionary pathways change in a longitudinal context?**

How can we accomplish this?

- Need to use multiple social surveys over multiple points in time
- Need to use administrative data to get access to qualifications and educational attainment data – social inequalities data

Why the UKHLS?

- A desire to study youth transitions in Britain
  - Traditionally this has been done with the use of the British Birth Cohort studies
  - There is a data black hole in the British dataset landscape from the 1970 British Birth Cohort Study to the Millenium Cohort Study in 2000
    - A 30-year gap (naturally nothing of interest happened in Britain from 1970-2000…)
  - The BHPS and UKHLS offer the ability to study this period of British history but there are several complications

Synthetic Birth Cohorts

- The first comes from the structure of the dataset itself
- The UKHLS is not a birth cohort study, it is a household panel study
- Synthetically derived cohorts are required for any analysis of cohort-based effects

Synthetic Birth Cohorts

- This is a relatively straight forward affair if using the main survey
  - Take birth year information from main survey wave and create cohort categories from them
- Focus of research is on specific age – the youth phase
  - Restricts our sample to individuals that have data when they were 16 years old.
- Focus of research is also not on birth years but school years cohorts
  - This means that we need to create our synthetic cohorts based on the active school year participation
    - Using birth day, month and year to calculate the school year allocation cohorts

Two Strategies

- Using the youth panel
  - Key identifiers: wHID (household), wPNO (person), and wYPWEGHT (weight)
  - Pooling youth panels together and then using birth information to create synthetic cohorts

  - For more see Gayle (2005)

Two Strategies - BHPS

- Using the main adult survey
  - Indresp + hhsamp
- Use day, month, year DOB
- Use Father and Mother unique id – bx_fnpid, bx_mnpid
- Full .do file to replicate:

https://github.com/Scott0atley/YouthTransitions/blob/main/data%20analysis/BHPS%20%2B%20UKHLS/1%20BHPS%20data%20set-up_UKHLS.do

Two Strategies - UKHLS

- Using the main adult survey
  - Indresp + hhsamp
- Use day, month, year DOB
- Create Father and Mother unique id
  - Loop through process of data cleaning
  - Rename pidp to father or mother pidp
  - Merge this with young person datafile from same family unit
- Check synthetic cohorts with NPD files using interview/birth dates recorded at admin level
- Full .do file to replicate:

https://github.com/Scott0atley/YouthTransitions/blob/main/data%20analysis/BHPS%20%2B%20UKHLS/2%20UKHLS%20data%20set-up_UKHLS.do

Admin Linkage

- Another issue comes from the type of school level qualifications data we need

- The BHPS has a breakdown of the type and grade of O'level/GCSE passes an individual attains
  - For the UKHLS the decision was made to shift this over to an administratively linked environment – using the National Pupil Database
    - (Ask me about how amazing this decision was later…)

Admin Linkage

- Administrative linkage to the NPD restricts our synthetic cohorts even further to only include UKHLS members that have valid records in the NPD database

- Those that take the UKHLS survey must agree to have their data linked

- NPD is slow – only has qualifications data up to ~2016

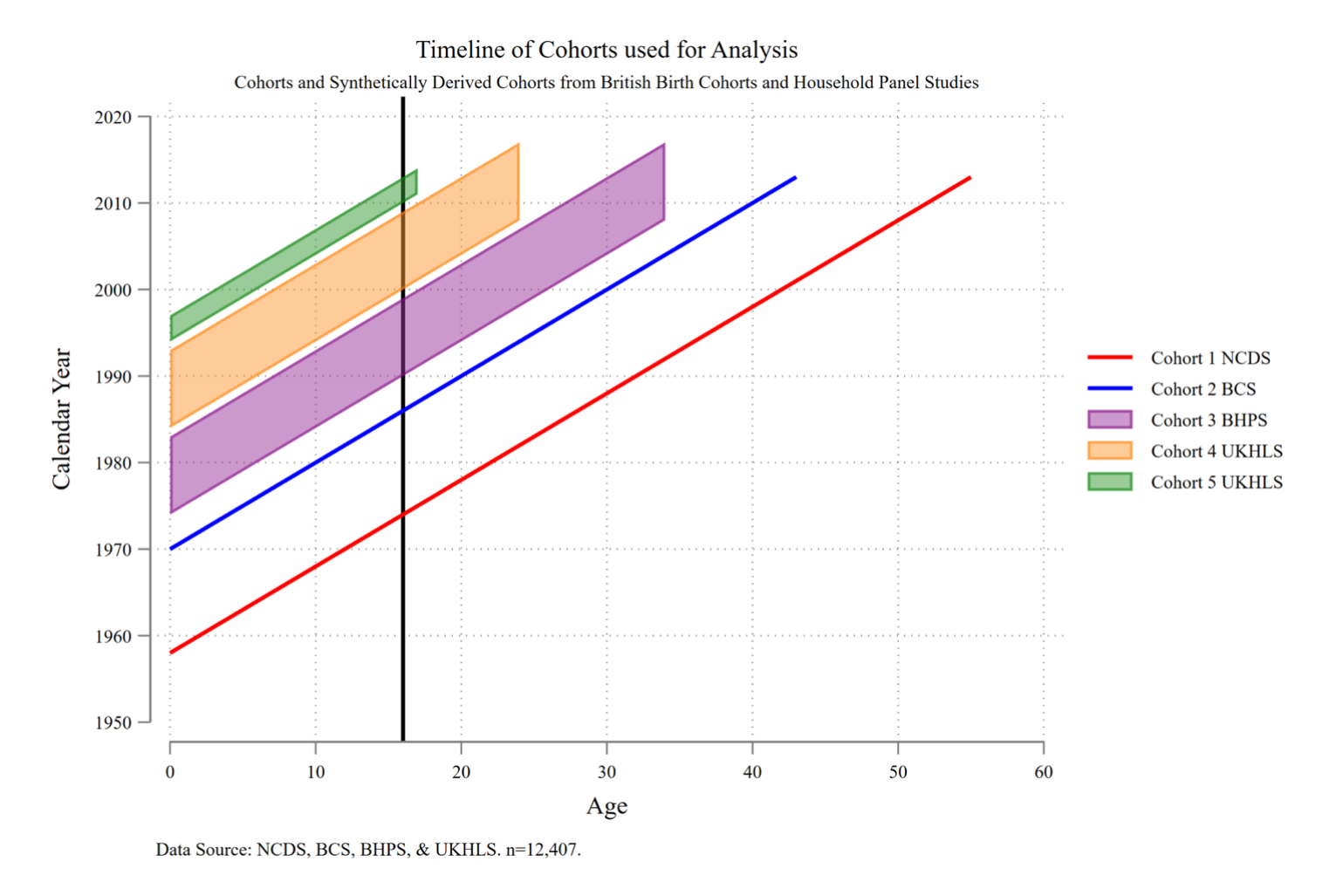
A Note on Administrative Data

- Delayed analysis by a good 6-9 months
- Lucky that the UKHLS + NPD linkage is well done, others are not as lucky
  - Look to MCS + NPD for examples of this
- In my view we should be pushing away from this reliance and linkage, let social surveys do what they do best
  - Others may have different views on this, welcome to that discussion!

School years

- Using birth day, month, and year data from the UKHLS requires secure access – as does using NPD data
- As we are restricted sample + valid NPD records to collect this data, yearly synthetic cohorts are not possible – some cohorts have n=20~
- Combining synthetic cohorts is required for statistical power
  - Five yearly and ten yearly cohorts attempted

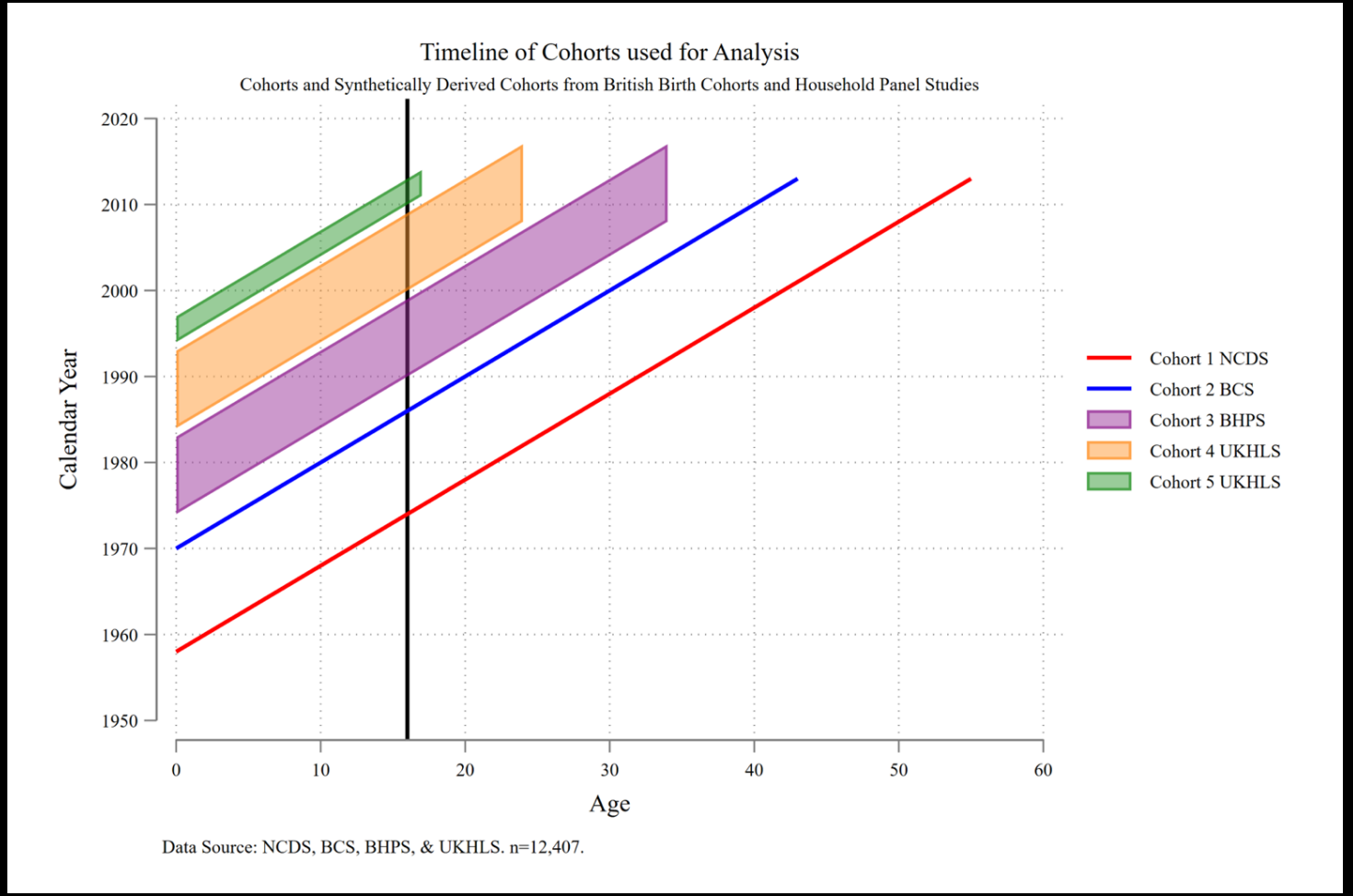
Result of Cohorts



Results of Cohorts

- We have the two older birth cohort samples
- We now also have three synthetic cohorts that use the household panel study
  - One uses BHPS sample and two use the UKHLS sample
- Functionally this means we can extend any analysis of youth transitions from a study period of 1974-86 to a period of 1974-2013*
  - Extension of 27 years

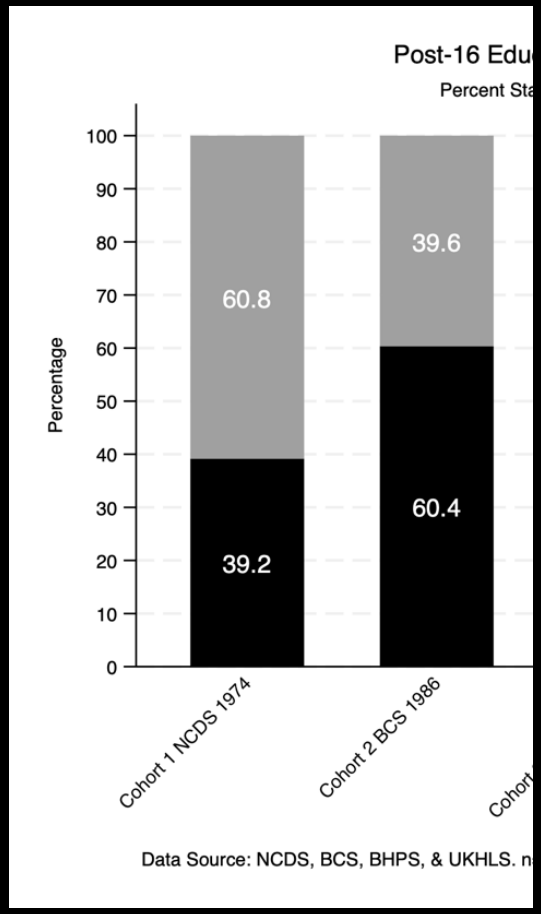
*(when youth are 16)



Full Cohort Sample

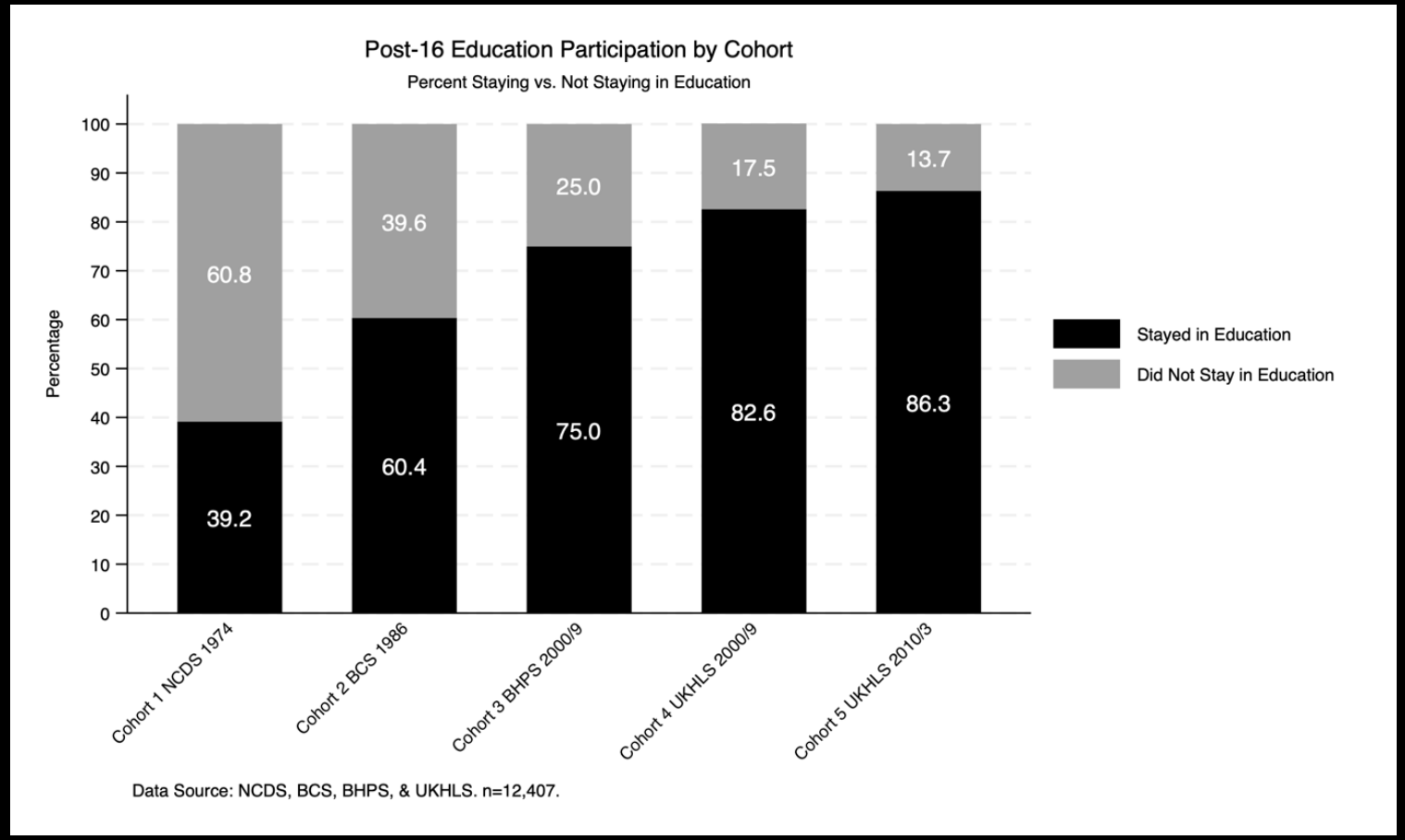
Cohort	Year	Sample
Cohort 1 NCDS	1974	8,411
Cohort 2 BCS	1986	1,574
Cohort 3 BHPS	1990/9	687
Cohort 4 UKHLS	2000/9	827
Cohort 5 UKHLS	2010/3	908
Pooled Cohorts	1974-2013	12,407

Story?


Post-16 Edu
Percent Sta
100
90
80
70
60
50
40
30
20
10
0
Percentage
60.8
39.2
39.6
60.4
Cohort 1 NCDS 1974
Cohort 2 BCS 1986
Cohort
Data Source: NCDS, BCS, BHPS, & UKHLS. n

Story

- The utility of the UKHLS is very apparent from this descriptive picture
- Using the UKHLS has extended our story of participation
- We are seeing an extreme sociological phenomena over the course of British history
  - Young people are staying in education in much larger quantities compared to the past



Social Inequality over time

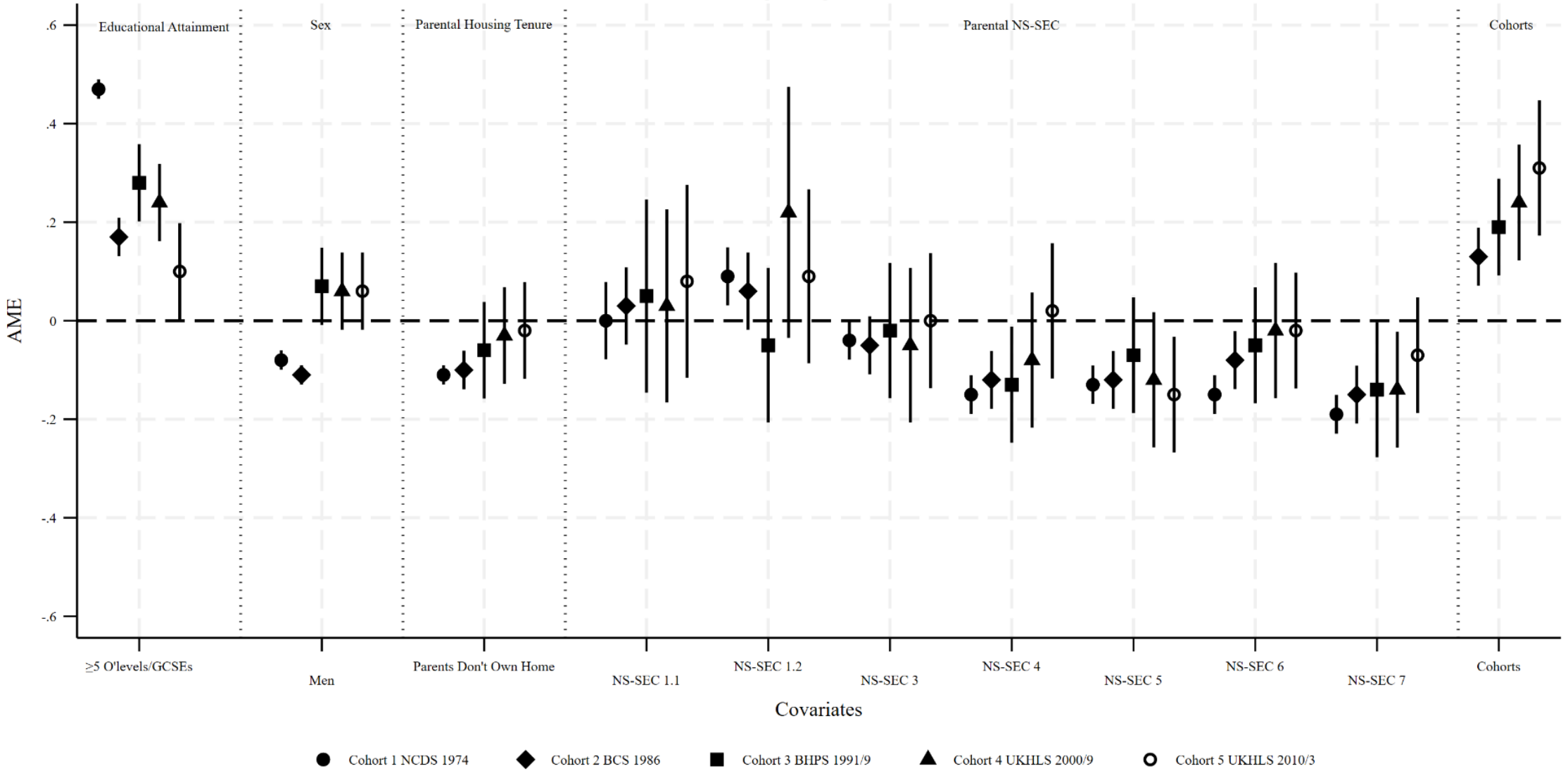
- To study this longitudinally, and to measure the effect of cohort level change we need to move to some more advanced modelling

- A logistic regression model with cohort level interactions with social inequality variables

- Social inequality variables:
  - Educational attainment = Binary of good O'level/GCSE passes or not
  - Sex = Binary of Male and Female
  - Housing Tenure = Parental Home Ownership or Not
  - NS-SEC = Parental Social Class with reference class NS-SEC 2
  - Cohort = Cohort level effect with reference Cohort 1 NCDS 1974

Impact of Covariates on Staying in Education

Average Marginal Effects



Data Source: NCDS, BCS, BHPS, & UKHLS. n=21,099. Adjusted for complex survey design. Reference Categories = < five O'levels/GCSEs, Male, Own Home, NS-SEC 2, Cohort 1 NCDS 1974.

BCS Sample Imputed.

Conclusions

- Marked decline in educational attainment, sex, housing tenure, and social class
- Marked rise in cohort level influences
- Death of social class? No. Death of the 16 year-old transition? Maybe
- Inequality is being pushed down the road

Questions