Advantages of Linked Administrative and Survey Data - Examples from Educational Research

LSE/NATCEN Conference, London
School of Economics, 24 April 2009

Lorraine Dearden
Director ADMIN
Institute of Education
Email: l.dearden@ioe.ac.uk
Introduction

- New linked admin/longitudinal data has potential to:
  - Get a better understanding of the implications of missing covariates in administrative (and possible survey data)
  - Get a better understanding of implications of attrition and non-response in survey data
  - Allow us to understand the implications and extent or recall bias in surveys……
  - Reduce the costs of longitudinal survey data
Illustrations from Educational Data linkage

- Already been extensive linkage of educational longitudinal survey data
  - MCS, ALSPAC, “Next Steps”
- This linkage has already started to be exploited
- But I think, from those concerned with methodological issues in empirical social research it has the potential to offer a lot more
Why?

- These are longitudinal panel data following children and families over time therefore shares all the advantages of longitudinal data that we are all familiar with

- BUT it is its linking to well established administrative education data which I think opens up a whole range of possibilities in advancing applied social research
Focus of this talk

- Concentrate on one of these data sets – Next Steps
- Follows a cohort of English children born in 1989/90 and has interviewed them and every year since they were 13/14 years old as well as their parents
- Sample originally drawn from National Pupil Database (NPD)
Next Steps link to NPD

- All individuals who participated in ‘Next Steps’ agreed to have their survey records linked to the NPD if in State Sector
- Means that alongside ‘Next Steps’ we have an administrative panel data set that covers every child in the state schooling system who are in the same school cohort
  - We know who the kids are in the NPD who are in Next Steps and we know who was invited to join Next steps but refused, later dropped out etc
Why is this exciting?

- NPD is increasingly being used to examine school policy and effectiveness but it is not clear how well it can do this as lacks detailed family background.

- Using ‘Next Steps’ we can assess how best to use administrative data to look at important and topical issues rather than just hoping that appropriate proxies are sufficient.
  - E.g. Contextualised Valued Added debate – how close is it really getting to measure what schools add?
  - Understanding sources of Ethnic inequalities in child outcomes
  - Does FSM really capture socio-economic status?
Panel Attrition and Missing Data

- Possibility of new ways of dealing with perennial problems associated with Longitudinal data
  - We know exactly how well children who drop out of Next Steps perform at school as we can follow them in the administrative data

- Admin data may be used to fill in gaps in survey data and vice versa
Examples of what we can do with ‘Next Steps’

- Measuring socio-economic disadvantage – are typical proxies in admin data good enough?
- Do School League tables (based on admin data) really do what they say on the tin?
- What is the best method for dealing with attrition and non-response in survey data?
Can educational admin data proxy social disadvantage?

- Huge debate about how good proxies in school administrative data are for social disadvantage
- Have good ethnicity measures, fsm status, but no information on parents employment, education or income/wealth
- Generally use postcode/local area information and/or FSM to proxy these missing covariates
- But how do they perform?
- Can assess this using NPD
Is FSM a good proxy of disadvantage?

- Construct a measure of socio-economic position using data on income, and occupation of parents

- This is what FSM should capture as only eligible if on benefits
OK but not great........

<table>
<thead>
<tr>
<th>Quintile of SEP</th>
<th>Not on FSM (%)</th>
<th>On FSM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Bottom)</td>
<td>9.78</td>
<td>10.21</td>
</tr>
<tr>
<td>2</td>
<td>17.32</td>
<td>2.66</td>
</tr>
<tr>
<td>3</td>
<td>19.39</td>
<td>0.61</td>
</tr>
<tr>
<td>4</td>
<td>19.70</td>
<td>0.31</td>
</tr>
<tr>
<td>5 (Top)</td>
<td>19.96</td>
<td>0.05</td>
</tr>
<tr>
<td>TOTAL</td>
<td>86.16</td>
<td>13.84</td>
</tr>
</tbody>
</table>
Ethnic Inequality in Child Outcomes

- Most of the evidence on ethnic inequality in child outcomes is based on administrative data
  - Sample sizes in most surveys too small to look at this issue properly
- Next Steps (and MCS) provide us with guidelines as to how to best use administrative data for all cohorts of children, to look at this important issue
- Can we proxy background characteristics just using admin data controls?
KS3 Results for Pakistani Males

<table>
<thead>
<tr>
<th></th>
<th>No Controls</th>
<th>NPD Controls</th>
<th>‘Next Steps’ Controls</th>
<th>NPD + ‘Next Steps’ Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPD sample</td>
<td>-0.314</td>
<td>-0.222</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSYPE sample</td>
<td>-0.356</td>
<td>-0.265</td>
<td>-0.129</td>
<td>-0.111</td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td>(0.070)</td>
<td>(0.068)</td>
<td>(0.069)</td>
</tr>
</tbody>
</table>

NB: Results show differences in standardized score outcomes
Not if interested in Ethnic Inequality

**Why?**

- Things like mother’s education differ hugely by ethnicity and typical proxies (area based census measures) can’t capture this
  - Could get census runs to overcome this potentially (education by ethnicity)
  - Could use other surveys to proxy this but not typically done
  - New linked survey data will allow us to come up with better proxies
Contextualized value added is the flavour of the month

Recognised the fact that traditional league tables did not take into account differences in prior academic achievement or socio-economic background
  - Hugely disadvantaged schools in poor areas even if adding a huge amount of value to the child’s education

CVA seen as solution to this
But….

- Can only control for variables in data
- What we see in Next Steps, however, is that things like home tutoring, time spent with child doing homework etc has significant impact on child outcomes
- Can’t control for this in Admin data so where it happens, school rather than parents get credit for this
- Implications for school league tables
  - Depends crucially whether this behaviour is a supplement or complement to the job being done by the school
- With Next Steps can come up with methods to limit these types of biases in CVA measures
Conclusions

- More and more survey data sets are going to be linked to admin data and this has huge potential
- Demonstrated some of the areas this should impact on in the education field – but easily transferable to other areas of social policy…