

# The Pros and Cons of Combining Birth Cohort Data

**This short SEED working paper is designed to highlight the value of and the pitfalls in combining and comparing data across large scale representative population cohorts. It was carried out as part of the Norface DIAL initiative by the SEED project (Social Inequality and its Effects on child Development: A study of birth cohorts in the UK, Germany and the Netherlands). In this working paper we draw on our expertise and experience in the research area of child development and cohorts we have utilised.**

## THE INCREASING AVAILABILITY OF BIRTH COHORT DATA

The development of large population and nationally representative birth cohorts has been going since their first introduction in the UK in the 1940s. They clearly provide a snapshot of what is going on in a country over a period of time and can provide insights into a wide range of societal issues, the potential mechanisms, and indicate where to target resources. They have become increasingly popular since the 1980s and they are a feature of the research landscape in many countries<sup>1</sup>. However, they are enormous undertakings with significant resources needed to plan, carry out, and maintain the data they generate. Therefore the

decision to start a cohort is not to be taken lightly. What we now have is a proliferation of studies many of which draw on the experience of those setting up earlier models in terms of sampling and the measures used. Cohort studies collect a range of rich data, but many have a primary focus such as, health data (ALSPAC/Generation R) or social and family data (MCS/NEPS/GUS/ELFE). With the increasing availability and accessibility of cohort data, it presents the possibility of using these cohorts to make comparisons within countries and/or between countries.

It should be noted that the use of multiple cohorts brings both opportunity and challenge. Working with cohort data requires a range of specialist knowledge such as, data management and statistical techniques, but this also creates the opportunity for multi-disciplinary and international collaboration of research<sup>2</sup>.

It is also important to acknowledge that the availability and accessibility of datasets varies. While some are accessible to researchers from all over the world (e.g., NEPS/MCS), others require the purchase of a license (e.g., ALSPAC/Generation R). Other data restrictions due to regulation can also

---

<sup>1</sup> Pearson, H. (2016). *The life project: The extraordinary story of 70,000 ordinary lives*. Catapult.

<sup>2</sup> Waldfogel, J. (2013). Socio-economic inequality in childhood and beyond: an overview of

challenges and findings from comparative analyses of cohort studies. *Longitudinal and Life Course Studies*, 4(3),

<https://dx.doi.org/10.14301%2Fllcs.v4i3.263>

affect the access and use of datasets. For example, using sensitive data within the MCS requires specialist training and access to a secure digital environment. Such restrictions can also limit the combining of datasets, meaning analyses must be run in parallel and results compared rather than simply running one analysis.

For access and international collaborations consideration about the language datasets and documentation is available in is also important. For some cohorts such as NEPS documentation is in both English and German but MCS/GUS and ELFE only have documentation (incl. questionnaires) in the host language (i.e. English and French respectively).

Increased availability of cohort data also opens the possibility of data linkage, particularly with administrative datasets, both to supplement the data but also for comparison. However this often brings us back to issues of data regulation and increased conditions of access.

### **DIFFERENT WAYS OF COMPARING AND COMBINING DATA**

Full harmonisation aims to bring the desired data together into a cohesive dataset(s), often requiring the re-formatting and structuring of variables so that they are as similar as possible. Parallel analyses is an easier approach as it doesn't require the harmonisation of each variable, but care must be taken that variables are still comparable. It is also important to note that harmonisation of

independent variables (for example maternal education or income) may be more straight forward than harmonisation of dependent variables, especially those which measure cognition or language. Measures commonly differ across cohorts (e.g., receptive vocabulary in NEPS vs. expressive vocabulary in MCS) and it is not always easy to establish how measures were selected or adapted for use in a given cohort.

In certain countries there are organisations that work towards creating data harmonisation. E.g., in the UK CLOSER<sup>3</sup> have funded work to harmonise certain variables, such as mental health and socio-economic measures, across UK cohorts.

### **WITHIN COUNTRY COMPARISONS**

This creates the potential for natural experiments, to capture and explore changes in specific aspects, and this is an effective approach if sampling and measures are comparable.

We have seen within country comparisons and the use of birth cohort data used to great effect within the UK, as an example. Comparisons between the different cohorts are commonplace in research to look at changes over long periods of time, but we also see policy and interventions being evaluated through the birth cohort data. For example, the Scottish Governments used two cohorts of the GUS study to test the effect of an early years programme related to children's exposure to a reading intervention.<sup>4</sup>

---

<sup>3</sup> <https://www.closer.ac.uk/>

<sup>4</sup> Bradshaw, P., King, T., Knudsen, L., Law, J. & Sharp, C. (2015) Early parent/child reading and its impact on language development and reading

enjoyment: comparing two Growing Up in Scotland cohorts *TECHNICAL ANNEX*. Edinburgh SCOTCEN

The timing of data collection and availability of cohort samples can also be of benefit to gauge changes within a country to global events. For example, in Ireland the timing of the GUI study made it possible to look at the effect of in-home childcare following the 2008 financial crisis. Similarly, in 2020 as the COVID-19 global pandemic took hold all participants in UK birth cohorts (NCDS/BCS70/MCS etc.) were sent a survey to rapidly collect data about the impact of the pandemic on people's lives. In the same vein, participants of all the starting cohorts of the NEPS were surveyed during the COVID-19 pandemic.

In terms of analysis, within country comparisons benefit from generally having a good level of harmonisation already with outcome measures and the age of assessments often being similar. Careful consideration about methodology and means of data collection are still required as the data remains subject to a great many caveats, the sampling, the weighting, and where measures are updated/re-standardised.

### **BETWEEN COUNTRY COMPARISONS**

As datasets become increasingly open access and available to researchers globally (UK data service/NEPS) it enables more comparisons between countries.

As highlighted earlier, access to certain datasets can be restricted, but such limitations create the opportunity for

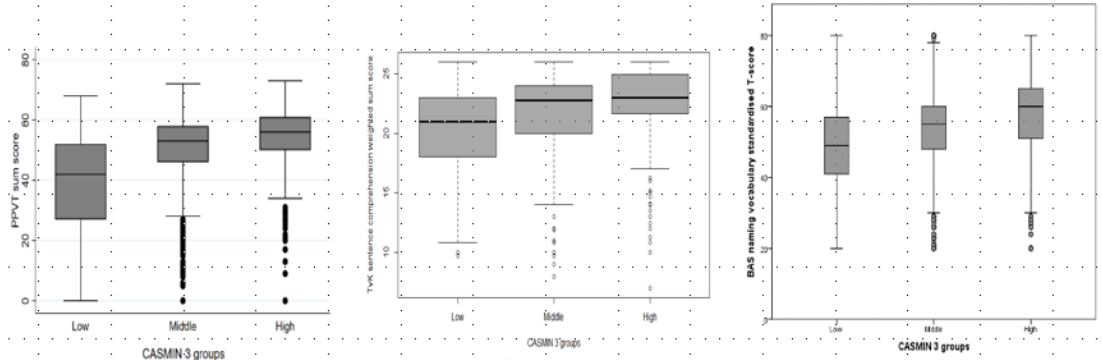
greater international collaboration and for methods of parallel analysis to be utilised.

Comparisons between countries enable us to explore and try to answer fundamental questions. For example, given similar economic circumstances measures and confounding factors would you expect the relationships between variables to hold? And if they don't, what population or policy differences are that that might be key to that relationship? The value of such comparisons is demonstrated by existing programmes such as PISA and UN SDGs and much is made of the results but explaining them is a complex activity.

An example of between country comparison from the SEED project as part of our own analysis was how you classify parental educational so that it can be used in a comparable manner across cohorts (NEPS/MCS/Generation R). We found the CASMIN framework made this possible as it had been applied to a number of countries (pre-dominantly European) including the ones in our project (Germany, UK, the Netherlands).

A strength of CASMIN is the number of educational categories within the framework and therefore the potential level of detail and comparison between different education levels. We found variation between the 3 datasets in the distribution and sample size across the full range of categories which led to categories being compressed.

Figure 1 Vocabulary and age 5/6 (start of school) across the three cohorts by mothers educational background at birth



Despite this reduction of categories, in Figure 1 we can see that the education gradient is apparent in all three cohorts although terciles 2 and 3 are closer and tercile 1 correspondingly lower in the Netherlands and Germany. We can also see that the language tests at five years function rather differently, with both floor and ceiling effects in both the latter cohorts.

Another example of a comprehensive comparison using cohort data across countries is the Too Many Children Left Behind<sup>5</sup>. Using data from the US, Canada, UK and Australia it identified differences across aspects of early child development, concluding that higher levels of social and emotional difficulties and behaviour problems were found in countries with higher levels of social inequalities (favouring Australia and Canada at the expense the US and the UK). While the

findings of this study are of great interest their global generalisation is limited as all data is from English-speaking, affluent, and Western countries. They were also initiated at the beginning of 2000s, whereas other more recent birth cohorts exist. When comparing cohorts the time in which data was collected is important to consider. For instance, when comparing MCS, ELFE, and NEPS SC1, the SES-patterns across countries in early care and education participation might be driven by or at least exacerbated by the fact that the children in the MCS were born a decade earlier than children in the NEPS SC1 and ELFE.

Other cross cohort comparisons have also been carried out more recently<sup>6 7</sup>.

<sup>5</sup> Bradbury, B., Corak, M., Waldfogel, J., & Washbrook, E. (2015). *Too many children left behind*. Russell Sage Foundation.

<sup>6</sup> Linberg, T., Schneider, T., Waldfogel, J., & Wang, Y. (2019). Socioeconomic status gaps in child cognitive development in Germany and in the United States. *Social Science Research*, 79, 1-31. <https://doi.org/10.1016/j.ssresearch.2018.11.002>

<sup>7</sup> Crosnoe, R. L., Johnston, C. A., & Cavanagh, S. E. (2021). Maternal education and early childhood education across affluent English-speaking countries. *International Journal of Behavioral Development*, <https://doi.org/10.1177/0165025421995915>

## COMMON CONSIDERATIONS WHEN LOOKING TO USE COHORT DATA

**The initial sample** – It is impossible for any sample and collection of data to be free from bias so it is crucial to understand and consider who is in the cohort sample. A common sampling issue is that those who are of a higher socio-economic background are more likely to engage and agree to participate in cohort studies. Some cohort samples can be large but are nationally representative, instead focusing on a particular region (Generation R/ALSPAC).

**Comparability of variables** – Demographic variables are available in all datasets but their harmonisation can also be challenging. For instance, with regard to income there are cohorts with pre-tax and those with post-tax income available.

Not many birth cohorts have cognitive outcomes in the age of 3 or 4 year old children (e.g., in Generation R and in ELFE, there is no language or math test in age 3 or 4). Comparability of so-called “non-cognitive” outcomes can also vary as they can stem from different informants (e.g., parents and/or teachers are asked to assess children’s behaviour or ability, as well as the children themselves).

The experience of education is also a key factor in many of these analysis but which varies considerably from country to country<sup>8</sup>.

**Attrition** – It is generally the case that participants who are lost between data collection sweeps or who withdraw from the study are not at random. Instead, they are more likely to be from lower socio-

economic backgrounds, and this must be considered in analysis as it creates a bias in the later collected data within the cohort.

**Weighting** – Cohort studies aim to achieve a representative sample such that any results are generalisable to the population. Due to recruitment and retention of particular groups therefore studies often oversample from those groups and thus require adjustment weights to be incorporated in the analysis (MCS) to balance the results for this. These procedures can be complex where certain groups are under-represented within the data or have a high attrition (MCS).

**Missing data** – Linked to the attrition of the cohort sample it is important to consider who, what and why data is missing such that the correct method can be applied for example only imputing based on the assumption of missing at random.

## MAIN MESSAGES

- Several cohorts that have been developed based on similar models and focus, but as yet none are identical and therefore some level of harmonisation is required when working with cohort data.
- Comparing children’s experiences across countries is potentially very meaningful especially when there is a distinct difference with relation to their experience of childcare policy.

---

<sup>8</sup> Schneider, S. L. (2021). The classification of education in surveys: a generalized framework for

ex-post harmonization. *Quality & Quantity*, <https://doi.org/10.1007/s11135-021-01101-1>

- The analysis of cohort data requires sophisticated theorisation and the use of complex statistical and analytical approaches. But they provide many opportunities for multidisciplinary research collaborations.

#### Cohorts referenced in this brief:

ALSPAC – Avon Longitudinal Study of Parents and Children [UK]  
<http://www.bristol.ac.uk/alspac/>

BCS70 – 1970 British Cohort Study [UK]  
<https://cls.ucl.ac.uk/cls-studies/1970-british-cohort-study/>

ELFE – French longitudinal study of children [France] <https://www.elfe-france.fr/en/>

GUI – Growing up in Ireland [Ireland]  
<https://www.growingup.ie/>

GUS – Growing up in Scotland [Scotland]  
<https://growingupinscotland.org.uk/>

MCS – Millennium Cohort Study [UK]  
<https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/>

NCDS – 1958 National Child Development Study [UK] <https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study-2/>

NEPS – National Education Panel Study [Germany] <https://www.neps-data.de/>

Generation R [Netherlands]  
<https://generationr.nl/>



This paper was developed as part of the work of the SEED Consortium. *SEED stands for Social InEquality and its Effects on child Development: A study of birth cohorts in the UK, Germany and the Netherlands* (Grant # 462-16-030). The consortium members are: Manja Attig, Gwendolin Blossfeld, Marie-Christine Franken, Wei Huang, Pauline Jansen, Claudia Karwath, Lisanne Labuschagne, James Law (PI), Cristina McKean, Robert Rush, Nathalie Tamayo Martinez, Hans-Günther Roßbach, Marc van der Schroeff, Jutta von Maurice, Helen Wareham, Anna Volodina and Sabine Weinert.

**Website:** <https://research.ncl.ac.uk/norfaceseed/> **Email:** [admin.seed@ncl.ac.uk](mailto:admin.seed@ncl.ac.uk)

This brief should be cited as: Law, J., Wareham, H., Volodina, A. Rush, R. (2021) The Pros and Cons of Combining Birth Cohort Data Publications Archive - DIAL ([dynamicsofinequality.org](http://dynamicsofinequality.org))