UCL Institute of Education

National Child Development Study

Age 11 Essays – Imagine you are 25, 1969

Alissa Goodman, Martina Narayanan, Stephen Brown & Tom Murphy

User guide to the data (First Edition)
December 2017

CENTRE FOR LONGITUDINAL STUDIES

Centre for Longitudinal Studies
Following lives from birth through the adult years
www.cls.ioe.ac.uk

CLS is an ESRC Resource Centre based at the UCL Institute of Education

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

NCDS Essays Project Team

The work of extracting, transcribing and documenting the essays was carried out by the following team based at CLS:

Research Team	Alissa Goodman, Martina Narayanan,
Transcription protocol, and supervision	Gemma Seabrook, Josie Orledge, Gearoid Garvey
Transcription of essays	Neville Thompson David Jessep Stephen Brown Aabid Ali
Data management	Tom Murphy

ACKNOWLEDGEMENTS

We would like to thank all the cohort members and their teachers, who generously gave their time to participate in this project and without whom this research resource would not now be available.

We would like to acknowledge the work of Virginia Morrow and Jane Elliott, previous colleagues at CLS who led a project to transcribe a subset of approximately 500 of the the age 11 essays that are also included in this deposit, with the support of funding from the Nuffield Foundation (reference SGS/32496).

Thanks to JD Carpentieri, Andrew Schwartz, Peggy Kern, and Benedetta Pongiglione who (together with the current authors) comprise the research team funded by ESRC to analyse these essays, and who have given their advice at various stages of the essays transcription. We are very grateful to all of those who supported the transcription work, including Gemma Seabrook, Josie Orledge, Gearoid Garvey at CLS, and to the transcription team, which included Neville Thompson, David Jessep, Aabid Ali, and Stephen Brown. Stephen Brown also supported the writing of this documentation.

We also wish to acknowledge the support of the Economic and Social Research Council (ESRC) who has funded the current transcription and research project (reference ES/N00650X/1). We also thank ESRC for its support of the National Child Development Study and the wider activities of the Centre for Longitudinal Studies, through the CLS Resource Centre 2015-2020 (reference ES/M001660/1) .

PREFACE

This document has been prepared to accompany the deposit, with the UK Data Service at the University of Essex, of the full available sample of essays written by cohort members from the National Child Development Study (NCDS), a continuing, multidisciplinary, national, longitudinal study.

The elements of the deposit, to which reference will be made throughout this document, are identified below. Users are advised that they will need to consult all elements of the documentation to gain a full understanding of the data.

NCDS Essays Deposit: Elements

Title	**Format**
NCDS essays	txt
NCDS essays accompanying dataset	SPSS/STATA/TAB
NCDS Age 11 Essays: Guide to the Dataset	PDF

CONTENTS

INTRODUCTION

This document has been prepared to accompany the deposit with the UK Data Service at the University of Essex, of all available essays collected when cohort members were eleven years old in 1969.

NCDS

The *National Child Development Study* (NCDS) originated in the *Perinatal Mortality Survey* (see SN 5565 on the UK data service), which examined social and obstetric factors associated with still birth and infant mortality among an initial 17,415 babies born in Britain in one week in March 1958. The study was augmented in subsequent childhood sweeps by immigrants to Great Britain born in the study's target week, bringing to the total NCDS sample to 18,558. Surviving members of this birth cohort have been surveyed on eight further occasions in order to monitor their changing health, education, social and economic circumstances – in 1965 at age 7, 1969 at age 11, 1974 at age 16 (the first three sweeps are also held under SN 5565), 1981 (age 23 – SN 5566), 1991 (age 33 – SN 5567), 1999/2000 (age 41/2 – SN 5578), 2004-2005 (age 46/47 – SN 5579), 2008-2009 (age 50 – SN 6137) and 2013 (age 55 – SN 7669). During 2002-2004, 9,340 NCDS cohort members participated in a bio-medical survey, carried out by qualified nurses (SN 5594, available under more restrictive Special Licence access conditions; see catalogue record for details).

There have also been surveys of sub-samples of the cohort, including in 1995 (age 37), when a 10% representative sub-sample was assessed for difficulties with basic skills (SN 4992), a Social Participation and Identity sub-study in 2007-2010 (age 49-52) in which a sub-sample of 220 were interviewed in depth about their lives and experiences of participating in the study (SN 6691). A subset of 500 age 11 essays that were transcribed in a previous project, and which are included in the current data deposit, are also separately deposited (SN 5790). Further NCDS data separate to the main surveys include a response and deaths dataset, public examinations records, parent migration studies, employment, activity and partnership histories - see the NCDS series page for details.

Further information about the NCDS can be found on the Centre for Longitudinal Studies website.

BACKGROUND TO THE ORIGINAL ESSAYS COLLECTION

When the children of the NCDS were eleven years old they were given a set of cognitive tests, together with a short questionnaire to complete at school about their interests outside school, the school subjects they enjoyed most, and what they thought they were most likely to do when they left secondary school. In addition, they were asked to write an essay about what they thought their life would be like at age 25. The instructions given were as follows:

> 'Imagine you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).'

Of the 15337 children who participated in the age 11 sweep of the NCDS, our records suggest that a total of 13732 (92.6%)[1][2] may have participated in the essay writing task about their imagined life at age 25. While a large majority of these original essays have been safely stored, unfortunately not all of them have survived. We provide details of the sample available for research further below.

Previous essay data deposits

While the essays were written in 1969, during the cohort members' final year in primary school[3], the current data deposit represents the first attempt to transcribe the full set of essays that were written and make them available digitally for large scale quantitative analysis.

At the time of data collection, it has been reported that there were no specific or immediate plans to analyse the essays written by the cohort members. Although some preliminary coding of the occupational aspirations of cohort members was carried out and archived together with the other quantitative data collected at age 11 (see variable n958 in SN 5565 - *National Child Development Study: Childhood Data, Sweeps 0-3, 1958-1974*), the essays were mainly seen as having scientific interest in the future once the occupational and family life trajectories of individual cohort members unfolded.

During the 1970s some analysis of a sub-sample of the essays was carried out which focused on the 'syntactic maturity' of the children in the study. A random sample of 521 of the total of 13732 essays were coded for their word count, which was included as a variable (n1857 – "2T Number of words in essay-sample only") in the main NCDS dataset for waves 0 to 3 which is available at the UKDS (SN 5565). However, users are advised to exercise caution in using this variable, as there are indications that it contains errors, but its accuracy can no longer be verified since the vast majority of essays from this sample were missing from the microfiche records held at CLS, and their paper test booklets destroyed. Please refer to the word count variable "WORDCOUNT" in the accompanying dataset of this deposit for the most accurate record of NCDS age 11 essay word counts.

In 2007, a sample of 495 essays was transcribed. In order to ensure sufficient numbers in subgroups to facilitate some simple comparisons to be made, a stratified sub-sample was extracted based on the three key variables gender, social class/family background and general cognitive ability.[4]

[1] This number is derived from the total number of study participants for whom an indication of participation can be found in any of the previous essay data deposits, or according to any derived essay-related variables held on record.

[2] Previous releases of the age 11 essays (e.g. SN5790) report that 13669 study members took part in the essay writing task. This total was based information contained in the variable n958 in SN5565 (which is described further below).

[3] Documentation to SN 5790 reports that the large majorty of essays were completed in April, May and June of 1969 when the cohort were in their final year of primary school. A small number (4%) were completed after the end of primary school, the very latest being in March 1970

[4] For further details, see J Elliott, V Morrow (2007) Imagining the Future: preliminary analysis of NCDS essays written by children at age 11 Centre for Longitudinal Studies Working Paper 2007/1, Institute of Education,

Derived variables from this subsample were coded and included whether or not the essay featured the the topics and themes that - according to the research team's assessment - emerged repeatedly in the children's essays. This enabled identification (for example) of how many of the essays included references to the cohort children's mother, father or siblings; how many discussed wanting to have children; and how many gave details about the skills they would use in their preferred occupation. The subsample of essays and derived variables were made available on the UKDS (SN 5790 - *National Child Development Study: Sample of Essays (Sweep 2, Age 11), 1969*). The essays from this sample have also been included in this release, and are flagged as "phase 1" in the "PHASE" variable on the accompanying dataset. The derived variables from SN 5790 are not included in this release.

In 2007-2010, a further set of 179 essays were transcribed and included in the following dataset: *Social Participation and Identity, 2007-2010* (SN 6691 at the UKDS). These essays were deposited together with a set of quantitative and detailed qualitative data. The essays included in this study have been excluded from this release, and are flagged as "phase 2" in the "PHASE" variable on the accompanying dataset. Researchers wishing to analyse these 179 cases are required to obtain SN 6691 under the more restrictive Special Licence access conditions under which that study is deposited.

Copies of the original essays (i.e. in the child's handwriting) have been stored on microfiche and archived at the Centre for Longitudinal Studies.

University of London, and J Elliott (2010) Imagining a gendered future: Children's essays from the National Child Development Study in 1969 Sociology 44 (6), 1073-1090.

THE TRANSCRIBED ESSAYS DEPOSITED IN THIS RELEASE

Transcription project

A major transcription project took place in 2016 and 2017 (referred to here as phase 3), in which all remaining essays available on the microfiche stored and archived at CLS were manually transcribed. With the exception of aforementioned 179 essays that are included in the Social Participation Study (transcription phase 2, SN 6691), this data deposit contains all NCDS age 11 essays that have been digitised, including both the set of essays transcribed as part of this project (phase 3), and additionally the sample of 495 essays (phase 1, SN 5790).

The aim of the transcription process was to reproduce the essays as accurately as possible, including spelling, grammatical and punctuation errors. For the current transcription phase (phase 3), if part of an individual word was illegible, and no reasonable guess could be made based on the context of the sentence, the word was transcribed to the fullest possible extent, with asterisks substituted for the illegible letters. If an individual word was deemed to be somewhat ambiguous due to poor legibility, the transcribers substituted their best guess based upon the context of the sentence, and marked the word with a single asterisk. A number of redactions of identifying material were also made, and are marked in the text between square brackets (e.g. “[NAME]”, [“PLACE”]).

As part of a data cleaning procedure, a small number of teacher’s notes were removed from the essays, and any duplicate essays identfied were checked against the microfiche records and removed or re-transribed as necessary.

Please see Appendix A for the complete guide to essay data entry, as originally provided to the transcribing team.

Please note that Illegible or partially illegible essays were treated slightly differently in the essays from the phase 1 transcription process (SN 5790). Here illegible text was replaced by the text “[Illegible]”, and the status variable on the accompanying dataset (called “ESSTAT” in this release) was coded to 2 (“illegible”) or 3 (“partially legible”). These two categories were not used in the phase 3 cases.

The achieved sample

While records suggest that 13732 cohort members participated in the essay writing task, the total number of essays deposited here is 10511. Of the remainder, there were 1447 cases where the microfiche appears to be present for that cohort member, but the essay page was either left blank or not included in the scan. There were 1535 cases where the microfiche was missing altogether. 225 cases are not included here because they formed part of the Social Participation Study (of which 179 of these are included as the phase 2 cases from SN 6691, which requires a special license), and 14 cases were classified as illegible/other missing during phases 1 or 2.

Table 1 (see appendix B) provides a profile of the current sample of 10511 transcribed essays, compared to two target populations. The first comparison group is the original complete NCDS sample (N = 18558). The second comparison group is the responders at age 11 (N = 15337). All three groups were compared based on a few central sociodemographic variables in order to examine different aspects of the representativeness of the the final sample of transcribed essays.

As can be seen in Table 1, there are a few significant differences between these samples, but in general the magnitude of those differences are small, even when they reach conventional levels for statistical significance. There are no differences in gender, mother’s age at birth, or parent’s education between the sample of cohort members with transcribed essays and the target populations. However cohort members from the final transcribed essay sample were from slightly higher father’s social class at birth, higher birthweight , and were of slightly higher general cognitive ability than both members of the original population, and of responders to the age 11 survey. This is consistent with known patterns of non-response at age 11 (noting non-response in childhood sweeps was in fact very low). It is also consistent with our presumption that some essays that are missing were due to non-completion by the cohort member (while for others, the completed essay became lost when transferred to the microfiche). There also are some national differences between the samples, with slightly lower than average representation in Wales and higher in Scotland. For ethnicity, there were no differences between the sample with completed/transcribed essays and the responders at age 11, but there were significantly fewer cohort members with non-caucasian ethnic backgrounds in the final essay sample compared to the complete NCDS sample, reflecting the fact that some of the non-British-born sample (known as the ‘immigrant boost’) were not added into the survey until the age 16 sweep. These kinds of missing data issues are very commonly observed in longitudinal studies and can be addressed and corrected for using analytic strategies such as multiple imputation and full information maximum likelihood approaches.

Future plans for deposit of further variables

At the time of deposit of this dataset, a research project is underway in which a number of derived variables are being created from the essays text which will be deposited in an updated version of this dataset in due course. The derived variables are being created using open content analysis, using machine learning tools. They will include topics (words or word-groups which cluster together) and other constructs such as those representing cognitive complexity. These will be fully documented and explained in the next data release.

ACCOMPANYING SPSS/STATA/TAB DATASET

An SPSS dataset to accompany to the NCDS essay text files has been supplied to the UK Data Service (also available in STATA or tab delimited formats) with the following variables:

Variable name	**Variable label**	**Additional Information**
NCDSID	NCDS serial number	The unique individual identifier for NCDS. This identifier also appears on other NCDS datasets available from the UK Data Service, and can be used to link the data longitudinally to other survey sweeps.
PHASE	Essay transcription phase	• Phase 1 = originally transcribed for SN 5790: National Child Development Study: Sample of Essays (Sweep 2, Age 11), 1969) • Phase 2 = originally transcribed for SN 6691: *Social Participation and Identity, 2007-2010* • Phase 3 = transcribed for the current project
ESSSTAT	Availability of transcribed essay	1 = coded 2 = illegible(phase 1 only) 3 = partially legible(phase 1 only) 4 = essay missing from microfiche 5 = microfiche missing 6 = other missing 7 = phase 2 essay not included here Categories 2 & 3 (Illegible and partially legible) are only applicable to essays from phase 1. For details on how illegible/partially legible essays were transcribed in phase 3, please see appendix A.
WORDCOUNT	Number of words in essay	

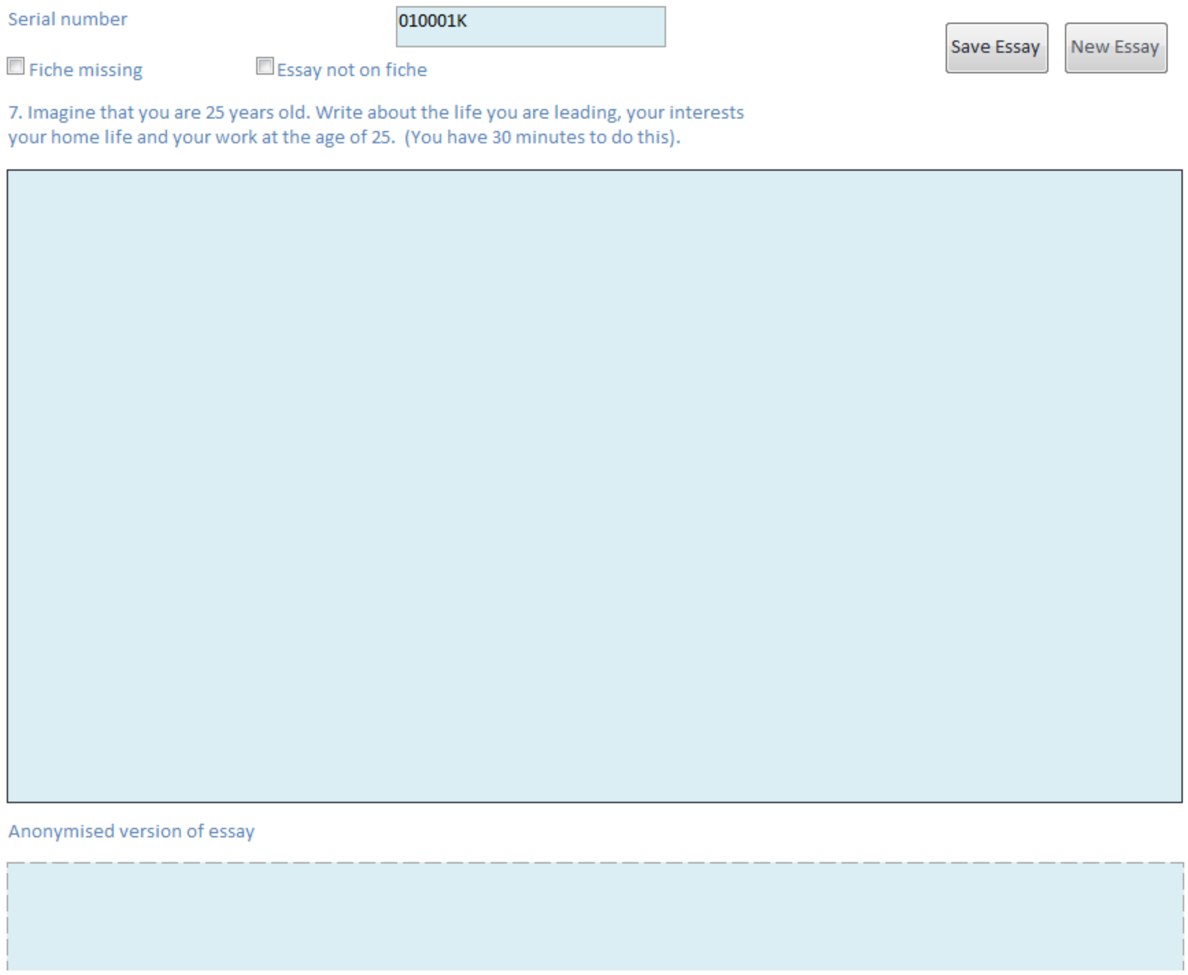
Appendix A: Instructions for essay data entry

Batches

All essays have been divided into Batches of 1500. One person should enter and anonymise each batch. It should then be passed to another person to check 10% of the entries.

Database

All essays should be entered into the databases produced for this task. The IDs are already present in the database and include all possible entries that should be made. If the microfiche cannot be found, the entry should be marked as such.



Serial number: 010001K

□ Fiche missing □ Essay not on fiche

7. Imagine that you are 25 years old. Write about the life you are leading, your interests your home life and your work at the age of 25. (You have 30 minutes to do this).

Anonymised version of essay

Microfiche

Essays can be found on the yellow topped fiche. A code is needed for entry. If the essay is present it will be the last item on the microfiche.

- If the fiche is missing tick Fiche missing
- If the fiche is found, but the essay is not on it tick Essay not on fiche

There is a comments field at the bottom where any comments can be made that might reflect other issues with the fiche or the entry. This is not a formal data field, but is more for reference if anyone were reviewing the entry.

Typing up the essay

The essay should be typed into the first box **exactly as it appears**. This means:

1. Reproduce all spelling mistakes and similar errors.
2. If a word might be wrong, place a * after it to indicate this e.g. football*
3. If a word cannot be read in full, use * to indicate illegible parts e.g. foot**ll
4. If a word is completely illegible use * to indicate the number of letters as best you can e.g. ********

Anonymising the essay
Once the essay is typed as is, copy it into the bottom field (Ctrl+C for copy, then Ctrl+V for paste are useful keyboard shortcuts that can be used in Access). You then need to remove any identifying information. The item removed should be replaced with a description of it in square brackets. For example for “My name is Bob” you would anonymise this to “My name is [name]. You should check and potentially remove:

- All names of the cohort member and their family/friends e.g. Uncle Bob, Bob who I am going to marry, etc. You can leave names of famous people if you are certain that’s who they are talking about (common examples might be footballers and pop stars).
- Names of small places. If a reference is made to a city e.g. London, Birmingham, Manchester, etc. this can be left as it is a large city that might have multiple cohort members (and attract additional cohort members). If a place is smaller and might have only one cohort member then it should be anonymised and replaced with [town], [village] or however the place is described on Wikipedia. The idea behind this is that the cohort member might be naming their home town and so it should be anonymised.
- Anything else you think might identify the cohort member. If they talk about something that might be unique enough to identify them, it should be removed. If in doubt, ask.

Checks
10% of all entries will be checked by another person for quality purposes. A small error rate is acceptable (and expected). This is to ensure that there is no significant error rate. After checking, feed back to the person whose work is being checked if any trends are noted so they can self-check for these in the future.

Appendix B: Representativeness of transcribed essays

Table 1: Comparison of final sample of transcribed essays to original NCDS sample and responders at age 11

	Transcribed essays N = 10511		Responders at age 11 N = 15337			Complete NCDS sample N = 18558		
	Mean	SD	Mean	SD	Diff	Mean	SD	Diff
Female gender	0.516	0.50	0.514	0.50		0.517	0.50	
Mother's age at birth (years)	27.5	5.7	27.5	5.7		27.5	5.7	
Parents' education (years)	11.3	1.8	11.3	1.8		11.3	1.8	
Social class birth manual	0.721	0.45	0.727	0.45	*	0.728	0.45	**
Birthweight (ounces)	117.8	18.3	117.6	18.4	*	116.2	20.5	***
Smoking during pregnancy	0.324	0.47	0.331	0.47	*	0.332	0.47	**
Country:								
England	0.848	0.359	0.844	0.363		0.844	0.363	
Wales	0.039	0.194	0.053	0.224	***	0.053	0.224	***
Scotland	0.114	0.318	0.103	0.304	***	0.103	0.304	***
General ability age 11 (score)	43.2	15.7	42.9	16.1	***	42.9	16.1	***
Euro-Caucasian ethnicity	0.971	0.17	0.972	0.17		0.959	0.17	***

Note: Differences between the sample of transcribed essays and the two target populations were tested using univariate logistic regression analyses.
Denoted significance level of the difference: * p < .05 ** p < .01 *** p < .001

Centre for Longitudinal Studies
UCL Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6860
Fax: 020 7612 6880
Email: clsfeedback@ucl.ac.uk
Web: www.cls.ioe.ac.uk