5

Technical report

FACS|AR

INSIGHTS, ANALYSIS AND RESEARCH

Pathways of Care Longitudinal Study: Outcomes of Children and Young People in Out-of-Home Care

Statistical Power, Selection Bias and Non-response Correction

Pathways of Care Longitudinal Study

Technical Report No. 5

Statistical Power, Selection Bias and Non-response Correction[1]

[1] This publication is to be read in conjunction with Steel, D. and Navin-Cristina, T. (2018). Initial Wave Weighting for the Pathways of Care Longitudinal Study. Technical Report Number 6. Sydney. NSW Department of Family and Community Services.

Prepared by
Fred Wulczyn, Linda Collins, Lijun Chen and Scott Huhr
Chapin Hall Centre for Children, University of Chicago

Published by
New South Wales Department of Family and Community Services (FACS)
Analysis and Research
223–239 Liverpool Road
Ashfield NSW 2131
Phone + 61 2 9716 2222

September 2017

ISBN: 978-0-6482696-5-6

Recommended citation
Wulczyn, F., Collins, L., Chen, L. and Huhr, S. (2017). Statistical Power, Selection Bias, and Non-response Correction. Pathways of Care Longitudinal Study: Outcomes of Children and Young People in Out-of-Home Care. Technical Report Number 5. Sydney. NSW Department of Family and Community Services.

Disclaimer
FACS funds and leads the Pathways of Care Longitudinal Study. The findings and views reported in this publication are those of the authors and may not reflect those of FACS. The authors are grateful for the reviewers' comments.

About the information in this report
All the information in this report is accurate as of May 2015. The analyses presented are based on a an almost final version of the Wave 1 unweighted data collected in face-to-face interviews with children, young people and caregivers; and FACS administrative data.

Pathways of Care Longitudinal Study Clearinghouse
All study publications including research reports, technical reports and briefs can be found on the study webpage www.community.nsw.gov.au/pathways

Study design by NSW Department of Family and Community Services (Analysis and Research); Australian Institute of Family Studies; Professor Judy Cashmore, University of Sydney; Professor Paul Delfabbro, University of Adelaide; Professor Ilan Katz, University of NSW; Dr Fred Wulczyn, Chapin Hall Center for Children University of Chicago.

Study data collection by I-view Social Research.

Advisors Expert advice and support has been provided by the CREATE Foundation; Aboriginal Child, Family and Community Care State Secretariat (AbSec); Connecting Carers and program areas.

Acknowledgements
We wish to extend our thanks to all the children, young people and caregivers who participated in interviews; childcare teachers, school teachers and caseworkers who participated in on-line survey interviews; and the data custodians in the relevant NSW and Commonwealth government departments. Ms Billy Black is a young person who grew up in care and designed the study artwork.

Ethics approval by University of NSW Human Research Ethics Committee (UNSW HREC) (Approval numbers HC 10335, HC 16542) and the Aboriginal Health and Medical Research Council (AH&MRC) of NSW Ethics Committee (Approval number 766/10).

Contents

List of figures

List of tables

1 Preface

Pathways of Care Longitudinal Study (POCLS) is funded and managed by the New South Wales Department of Family and Community Services (FACS). It is the first large-scale prospective longitudinal study of children and young people in out-of-home care (OOHC) in Australia. Information on safety, permanency and wellbeing is being collected from various sources. The child developmental domains of interest are physical health, socio-emotional wellbeing and cognitive/learning ability.

The overall aim of this study is to collect detailed information about the life course development of children who enter OOHC for the first time and the factors that influence their development. The POCLS objectives are to:

- describe the characteristics, child protection history, development and wellbeing of children and young people at the time they enter OOHC for the first time
- describe the services, interventions and pathways for children and young people in OOHC, post restoration, post adoption and on leaving care at 18 years
- describe children's and young people's experiences while growing up in OOHC, post restoration, post adoption and on leaving care at 18 years
- understand the factors that influence the outcomes for children and young people who grow up in OOHC, are restored home, are adopted or leave care at 18 years
- inform policy and practice to strengthen the OOHC service system in NSW to improve the outcomes for children and young people in OOHC.

The POCLS is the first study to link data on children's child protection backgrounds, OOHC placements, health, education and offending held by multiple government agencies; and match it to first hand accounts from children, caregivers, caseworkers and teachers. The POCLS database will allow researchers to track children's trajectories and experiences from birth.

The population cohort is a census of all children and young people who entered OOHC for the first time in NSW between May 2010 and October 2011 (18 months) (n=4,126). A subset of those children and young people who went on to receive final Children's Court care and protection orders by April 2013 (2,828) were eligible to participate in the study. For more information about the study please visit the study webpage www.community.nsw.gov.au/pathways.

2 Executive Summary

In contrast to research that relies exclusively on administrative records that cover an entire service population, surveys that involve a sample of service recipients have to contend with issues of statistical power, non-response bias, and survey weights if and when survey respondents differ from non-respondents. For the Pathways of Care Longitudinal Study (POCLS) each of these issues is relevant given how the study is being conducted. Statistical power addresses the question of sample size and whether the null hypothesis will be rejected given that it is in fact false (Cohen, 1992; Faul, Erdfelder, Lang, & Buchner, 2007). During the initial design discussions, when drawing a sample from a study eligible population was under consideration, the POCLS research team wanted to understand how many children would be needed to detect effects (i.e., differences in outcomes among groups of children) of a moderate size. In the end, the design included all children deemed to be study eligible, with study eligibility determined by whether a final care and protection order assigning custody of the child to the Minister had been issued by the court. The study of statistical power was used to shape those decisions. The study of non-response bias was carried out in order to contend with the fact that only some members of the study eligible population actually participated in the survey. The statistical differences that emerged suggest that a selection process was at work. Because respondents differ from non-respondents in ways that may also be related to the outcomes of interest (e.g., how long children stay in care, placement stability, and child wellbeing), generalisations from the POCLS sample to the larger study eligible population have to be approached cautiously. Weighting provides a way to balance the respondent sample so that it more accurately reflects the larger, study eligible population.

Each of these issues is taken up in this report. Although the sections stand alone, they are presented as a single, integrated technical document because together they represent a framework for understanding the POCLS sample and how it can be used. The technical report starts with the power analysis. The approach taken focuses on a statistical model of language development over time, with an emphasis on average scores at Wave 1 and change in scores across subsequent waves. The second section examines the ways in which respondents differ from non-respondents. The discussion is framed around the issue of selection bias. In the POCLS, respondents volunteered to participate. Thus, differences between respondents and non-respondents are the result of a self-selection process. In turn, the evidence of non-response bias points to a need to correct for the selection process, which is the subject of the third section. Using a propensity score framework, the empirical justification is provided for applying propensity weights to correct for the non-response bias in the sample.

3 Power Analysis for the Pathways of Care Longitudinal Study

3.1 Introduction

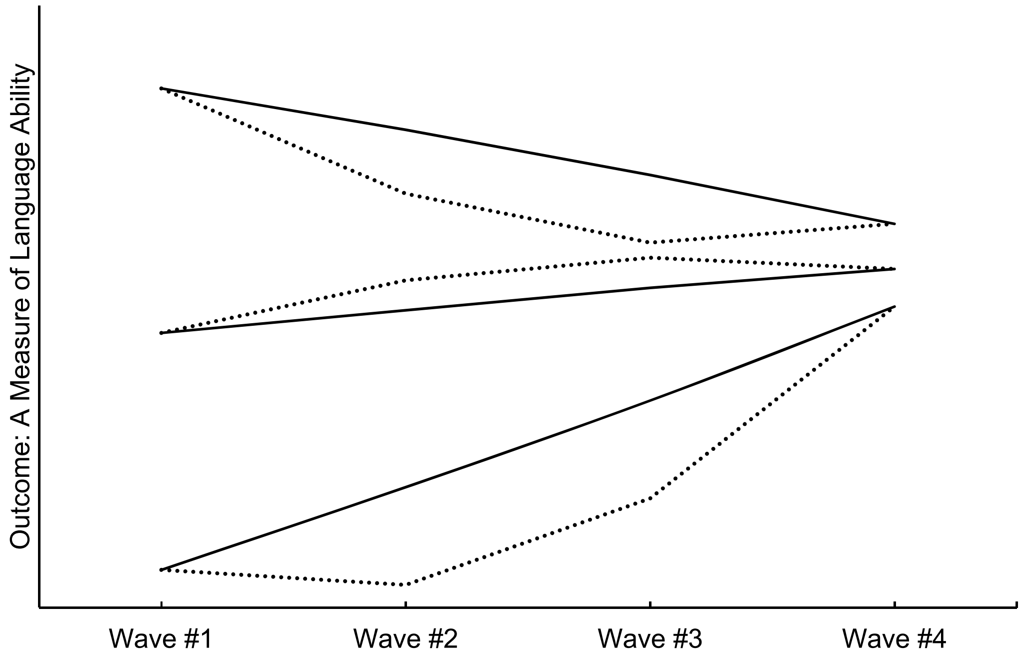
At its most basic level, the POCLS is about children in out-of-home care and the factors that affect their growth and development. When children are placed in out-of-home care, the public child welfare agency assumes responsibility for their care and protection. For the children who are placed on a final care and protection order, the public agency has to make sure children receive the health, behavioural health, educational services, and other life experiences they need to thrive over their life course, just as any parent would.

Observational studies of growth and development involving children in out-of-home care are necessarily challenging. Although the basic idea is to understand changes in wellbeing and the impact out-of-home care has on wellbeing, children differ with respect to when, developmentally speaking, the first out-of-home placement takes place. These initial differences are a function of age and early life experiences. From the POCLS perspective, the initial differences also affect the rate at which a child's growth and development will likely unfold. In other words, developmental differences observed at the outset (i.e., Wave 1) can have their own independent effect on how children grow and develop.

To highlight these issues within the context of a power analysis, Figure 1 represents the outcomes from a hypothetical measure of a child's language ability (the y-axis) measured over time (the x-axis). The figure illustrates different features of the pathways that outcomes may follow from wave to wave. For the purposes of this illustration, assume that the 'children' represented in Figure 1 are the same age and at the same educational level at Wave 1. The features accented in Figure 1 are as follows.

- Children will enter the study with different levels of language ability (represented by the dot at Wave 1). In the context of a regression model, the Wave 1 measure of language ability may be regarded as the intercept (or α).
- Language ability may rise or fall at some rate over time (represented by the straight lines). Change in language ability with respect to time is the slope (or β_1).
- Change in language ability may be faster initially and then slower later on (or vice versa). Changes in the rate of change, which are represented by the curves and typically captured with quadratic term in the regression model, are also associated with a slope (or β_2).

Figure 1: Features That Growth Patterns for a Measure of Language Ability Might Follow Over Time.



In practice, each child's path will likely exhibit one or more aspects of each of these three features. The patterns will also vary depending on the child's characteristics and experiences in placement. Estimation of parameters representing these features and the effect of child characteristics on these features will be a part of any reasonable modelling approach for longitudinal data from the POCLS.

3.2 Proposed Approach

The data forming the basis for the power study come from the National Survey of Child and Adolescent Wellbeing (NSCAW). This is a United States nationally representative survey of over 5,500 children from nearly 100 counties or child welfare agencies in 38 states who were investigated for child abuse and neglect within a 15-month period starting October 1999. The power study is based on a sub-sample of infants who were in out-of-home care and aged 9 to 15 months at their Wave 1 assessment – much like what we would expect when analysing a sub-sample of infant entrants in the POCLS data. In the NSCAW, these infants were followed up on three more occasions after the baseline survey: at median observation times of 15, 31, and 58 months after the Wave 1 survey. All of the three-wave power analyses referred to in this report are based on data from Waves 1, 3, and 4.

From the NSCAW data, we chose the outcome variable of interest to be a child's normalised score on the Pre-school Language Skills (PLS) instrument.

Four child characteristics were found to be related to the PLS score and were simply coded as 1 (present) or 0 (not present).

- The child was physically or emotionally abused (or not).
- The child was in a kinship care placement at Wave 1 (or not).
- The child was identified as white (or not).
- The child was in an urban setting (or not).

Using these data, the original question of power translates to: How many children are needed in the data so that abuse, care type, ethnicity, and setting may be simultaneously detected as significantly related to observed changes in children's PLS scores over time?

3.3 Complexities

The best prospective power analyses use:

1. A statistical model and data analysis process as similar as possible to that expected for the study data,
2. Data-based estimates for model parameters, and
3. An unbiased method for determining the statistical significance of model characteristics of interest.

For the POCLS, item 1 is complex because of the nested nature of the data. Each child will have multiple measurements (Waves 1, 2, 3, and 4) and children are served within potentially very different units (e.g., a non-governmental organisation (NGO) or District Office). Fortunately, the NSCAW data have a similar nested structure, which satisfies item 1.

Although an appropriate comparison to the POCLS data, the NSCAW data pose particular difficulties for item 2 since the data come from a cluster sample – primary sampling units (PSU) were chosen first, then agencies within those PSUs, and finally, children were chosen from within the agencies. Observations need to be appropriately weighted in any data analysis and the appropriate use of weights in data analysis of a nested statistical model is a matter still under some discussion in the literature. In the power analysis reported here, we implemented the method described by Pfeffermann and colleagues (Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998) along with advice specific to the NSCAW provided by Christ, 2007.

For item 3, the statistical tests of significance for model parameters or sets of model parameters can vary among statistical software programs since no general solution to this problem currently exists (Faraway, 2006). Our approach was to use a straightforward likelihood ratio test (LRT) to determine the p-value for the test of whether a model including certain features and/or child characteristics is a significant improvement over a model without them.

The p-value is the probability that we would incorrectly determine that the models are significantly different if, in fact, they are not. We require the p-value to be small: below $\alpha=0.05$ in Table 2.1 and below $\alpha=0.10$ in Table 2.2.

The LRT will tend to underestimate the p-value when sample sizes are small (Faraway, 2010, p. 158). That is, the test can sometimes overstate the importance of the effects of features and/or child characteristics in the model. Faraway (Faraway, 2006; 2014) suggests calculating the p-value for the LRT using a parametric bootstrap method. We did not implement this suggestion in the power study because this would mean that for each of the 500 simulated data sets developed for the power study, we would need to parametrically bootstrap, say, 200 times from the fit of the model to the simulated data. That approach was not feasible, so we determined the p-value directly from the LRT without bootstrapping.

We did investigate the risk of underestimating the p-value from the LRT and determined that with sample sizes of 200 children and above, this was not an issue. Tables 2.1 and 2.2 only consider sample sizes at least this large. When the eventual POCLS data are analysed for research, there will be just one data set, not 500 simulated data sets. We do suggest determining p-values using parametric bootstrapping of the LRT for analysis of the observed POCLS data.

In practice, model comparison during data analysis is typically accomplished by comparison of criteria developed to balance the improved model fit always experienced when more features or child characteristics are included vs. the additional complexity of a larger model. Common criteria of this type are the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). These criteria do not provide a p-value and are thus not useful for a power analysis study, but will be important for data analysis of the observed POCLS data.

3.4 Results

For our sample size study (i.e., power analysis), we estimated parameters from existing data with similarities to the anticipated POCLS data. As noted, from the parameter estimates and the model, we simulated 500 data sets ‘like’ the existing data (except with more children). For each simulated data set, we compared the fit of models with and without certain features and/or child characteristics. The proportion of simulations where the model with the features included is significantly different from the model without provides an estimate of power to detect these features and/or the effects of child characteristics. This is a parametric bootstrap estimate of power (Efron & Tibshirani, 1993).

Tables 2.1 and 2.2 provide the results of the power study[2]. A power level of 80% or higher indicates very good probability of detecting a feature. The tables indicate where power estimates are at least 60%, 70%, 75%, 80%, or even above 90%. The tables make clear that at any reasonable sample size we do not have power to detect a significant effect from all four child characteristics (variables) simultaneously on one feature alone (either the starting level, linear growth, or changes in growth rate over time – a curve). However, with four waves of data and 350 or more children in the analysis, we have good power to find that all four variables contribute simultaneously to at least one of the three growth-pattern features.

Table 2.1: Power Analysis With α=0.05

			(4) Number of Children in Simulated Data Sets						
(1) Feature(s) to Detect	(2) Least No. of Variables Significant	(3) No. of Waves	200	250	300	350	400	450	500
Start	2	3	80%+	90%+	90%+	90%+	90%+	90%+	90%+
Start	2	4	60%+	80%+	90%+	90%+	90%+	90%+	90%+
Start	3	3				60%+	70%+	80%+	80%+
Start	3	4				70%+	75%+	80%+	90%+
Start	4	3							
Start	4	4							
Line	2	3	80%+	80%+	80%+	80%+	80%+	80%+	80%+
Line	2	4	90%+	90%+	90%+	90%+	90%+	90%+	90%+
Line	3	3							
Line	3	4	70%+	70%+	70%+	70%+	75%+	60%+	70%+
Line	4	3							
Line	4	4							
Curve	2	3						60%+	60%+
Curve	2	4				60%+	60%+	70%+	75%+
Curve	3	3							
Curve	3	4							
Curve	4	3							
Curve	4	4							
At Least 1	2	3	90%+	90%+	90%+	90%+	90%+	90%+	90%+
At Least 1	2	4	90%+	90%+	90%+	90%+	90%+	90%+	90%+
At Least 1	3	3	75%+	80%+	80%+	90%+	90%+	90%+	90%+
At Least 1	3	4	80%+	90%+	90%+	90%+	90%+	90%+	90%+
At Least 1	4	3							60%+
At Least 1	4	4				60%+	70%+	60%+	75%+

[2] It should be noted that the power results are determined by the size of effects in the NSCAW data.

Further, at any anticipated sample size from a homogeneous subset of the POCLS data, it seems we should easily be able to detect contributions from two variables on either the starting value or the rate of growth. However, it is more difficult to detect changes in the rate of growth (a curve) and this is where we again see the advantage of having four waves of data rather than just three.

Estimated power to detect feature(s) in the pattern of children's language ability is listed by (1) feature (starting ability, rate of change (line), changes in the rate of change (curve), or at least one of these three features; (2) how many variables among four are required to be significant predictors of language ability for the feature; (3) how many waves of data are used in the analysis of the data (three or four); and (4) how many children are in the study. The attrition rate is 15% at each wave of data collection in the simulated data. The 'level of significance' required for any two models to be declared significantly different is α=0.05.

Table 2.2: Power Analysis With α=0.10

			(4) Number of Children in Simulated Data Sets						
(1) Feature(s)	(2) Least No. of Variables Significant	(3) No. of Waves	200	250	300	350	400	450	500
Start	2	3	90%+	90%+	90%+	90%+	90%+	90%+	90%+
Start	2	4	80%+	90%+	90%+	90%+	90%+	90%+	90%+
Start	3	3			60%+	70%+	80%+	80%+	80%+
Start	3	4			60%+	75%+	80%+	80%+	90%+
Start	4	3							
Start	4	4							
Line	2	3	80%+	90%+	80%+	90%+	90%+	90%+	90%+
Line	2	4	90%+	90%+	90%+	90%+	90%+	90%+	90%+
Line	3	3							
Line	3	4	80%+	80%+	80%+	80%+	80%+	80%+	80%+
Line	4	3							
Line	4	4							
Curve	2	3			60%+	70%+	60%+	70%+	75%+
Curve	2	4			70%+	75%+	70%+	75%+	80%+
Curve	3	3							
Curve	3	4							
Curve	4	3							
Curve	4	4							
At Least 1	2	3	90%+	90%+	90%+	90%+	90%+	90%+	90%+
At Least 1	2	4	90%+	90%+	90%+	90%+	90%+	90%+	90%+
At Least 1	3	3	80%+	80%+	90%+	90%+	90%+	90%+	90%+
At Least 1	3	4	90%+	90%+	90%+	90%+	90%+	90%+	90%+
At Least 1	4	3					60%+	60%+	60%+
At Least 1	4	4			60%+	70%+	75%+	80%+	80%+

This table is derived in the same way as Table 2.1, except that α=0.10. So, it is easier to declare a feature as significant, which leads to higher estimates for power. In practice, peer-reviewed social science research is regularly published highlighting results of interest with such significance levels. Thus, we also studied the power available at several sample sizes under this slightly relaxed condition.

3.5 Summary

The power analysis considered samples ranging in size from 200 to 500 children. In addition, we tested models focused on differences in language development at Wave 1 (start), changes in language development over time (line), and changes in the rate of change (curve). The results suggest samples of 200 children or more are adequate for detecting significant effects (using either size $\alpha=0.05$ or 0.10) when the analysis considers differences in outcome language skills in this example at Wave 1, with a two-variable model. Linear effects that capture the change over time in language skills are detectable with samples of 200 children or more and three or four waves of data. Three variable models require larger sample size; four variable models require samples of more than 500 children. Models that examine changes in the rate of change (curve) require larger sample sizes. Models with three or four variables require large samples in order to detect significant differences in the rate of change. If the expected significance level is relaxed to 0.10, the basic conclusions are, for all intents and purposes, unchanged. Larger number of variables and more complex outcome structure naturally require larger samples.

4 Selection/Non-Response Bias in the POCLS Sample

4.1 Introduction

Originally conceived during the 2006-2008 period, the POCLS is designed to deepen what FACS knows about the wellbeing of children in out-of-home care and the factors that influence their outcomes. The study design calls for children placed on final care and protection orders to be followed over roughly five years and four waves of data collection. Data collection links administrative data with data collected from carers, childcare workers or teachers, caseworkers and the children/young people themselves.

To be considered eligible for the study, a child or young person would have had to enter out-of-home care for the first time between May of 2010 and October of 2011. This group of children is known as the population cohort. Of children in that group (N = 4,126), children who were placed on a final care and protection order were then considered study eligible (N = 2,827). From the final care and protection orders (final orders)/study eligible cohort, FACS recruited caregivers to participate in an interview (N = 1,788). Of that group, 1,285 completed the Wave 1 interview[3].

In this study of non-response in the POCLS, we examine two key points in the process of selection into the study: selection from the population cohort into the final orders/study eligible cohort and from the final orders/study eligible cohort into the group of children/young people for whom an interview with carers was completed. With respect to the former, selection into the final orders/study eligible cohort speaks to differential experiences in out-of-home care. In NSW, a significant portion of children entering out-of-home care leave after a relatively brief placement. By conditioning study eligibility on whether the child was placed on a final care and protection order, the POCLS focuses on children who have been and likely will be in care for some time. Selection from the final orders/study eligible cohort into the group of children and young people whose carers were interviewed speaks to a slightly different issue. In surveys of this type, for reasons having to do with caregiver willingness to participate, trouble coordinating interview schedules, and other reasons, one cannot expect a 100% response rate. For this reason, it is important to understand who completed interviews in the likely event interviews involved a non-random subset of the final orders/study eligible cohort, which could

[3] We could have but did not, for this round of the analysis, consider if and how the population of study eligible children differed from those children whose caregivers agreed to be recruited. Children restored at the time of the Wave 1 interview were not invited to participate in the Wave 1 interview for practical and ethical reasons but joined the study at Wave 2. Please note that our analysis has excluded cases without identified service districts or with ‘statewide services’. We also excluded cases with only respite care episodes when creating the foster care spell data file. Thus the total number of cases is reduced by 80 to 4,046 cases.

influence how the study findings generalise to the larger population of study eligible children.

4.2 Study Eligibility

Children placed between May of 2010 and October of 2011 were candidates for study eligibility once they received final orders. Final eligibility for the study was based on whether a final care and protection order transferring parental responsibility to the Minister was entered with the Children's Court. As a general matter, final care and protection orders are correlated with length of stay because final orders are issued after a judgment about restoration has been made. Children with final care and protection orders may yet be returned to their parents, but the likelihood of restoration goes down with the issuance of the final order. The POCLS focuses on children with final orders in order to better understand what happens developmentally to children for whom FACS has assumed long-term responsibility.

Table 3.1 presents select characteristics of the study population by final order/study eligibility status. Caregivers of study eligible children were later invited to an interview. Overall, 70% of the children admitted during the study window became eligible (e.g., received final orders).
As expected, final orders/study eligibility and length of time in out-of-home care are inversely related. Among children in care for less than one month, only 21% received a permanent care and protection order. In contrast, 87% of the children in care for 24 months or more were study eligible.

Males were slightly more likely to receive final orders after admission than females. Final orders were much higher among young children – 89% of all infants were eventually eligible for the study, whereas only 29% of the 13 to 17 years became eligible.

By and large, Aboriginal children were about as likely to receive final orders and thus be eligible for the study as were non-Aboriginal children. Children of Torres Strait Islander descent were more likely to be eligible but there were very few Torres Strait Islanders in the population cohort; for convenience of analysis, they are grouped together with other Aboriginal children.
Most children placed in NSW are placed in family settings including both foster and kinship homes. Eligibility tended to lower among children placed in non-family settings (e.g., residential care). Among children placed with kin, final orders were less common among children placed with Aboriginal kin (62%).

With regard to maltreatment history prior to entry into out-of-home care, study eligibility does align with the number of ROSH reports (Risk of Significant Harm). Children with no ROSH reports were the least likely to receive final orders (37%), whereas children with the highest number of ROSH reports were the most likely to become study eligible (79%). Similarly, the number of substantiated maltreatment reports was correlated with receiving final orders/study eligibility. Only one-half of the children with no history of substantiated reports were included in the final orders/study eligible sample. Three quarters of the children with one or more substantiated allegations were study eligible. Of the predominant maltreatment issues reported, those with either neglect, physical abuse, or sexual abuse issues were more likely to receive final orders (76%) than maltreatment issues involving emotional abuse or mixed maltreatment issues. However, children with no reported maltreatment issues are about as likely to become study eligible (66%) as the latter groups of children.

Finally, receiving final orders/study eligibility varied significantly by District Office. Among District Offices with 50 or more admissions between May 2010 and October 2011, final orders/study eligibility rates ranged from 51% to 77%.

Table 3.1: Pathways of Care Longitudinal Study Sample by Study Eligibility Status and Selected Child Characteristics

Characteristics	Eligible	Not Eligible	Total	Eligible	Not Eligible	Total
Total	2,817	1,229	4,046	70%	30%	100%
Time in Care						
Less than 1 month	140	514	654	21%	79%	100%
1 to 2 months	49	79	128	38%	62%	100%
2 to 3 months	30	64	94	32%	68%	100%
4 to 6 months	120	95	215	56%	44%	100%
7 to 12 months	279	96	375	74%	26%	100%
13 to 24 months	258	93	351	74%	26%	100%
Over 24 months	1,941	288	2,229	87%	13%	100%
Gender						
Female	1,369	662	2,031	67%	33%	100%
Male	1,448	567	2,015	72%	28%	100%
Age at Placement						
Infants	881	112	993	89%	11%	100%
1-5 years	1,024	330	1,354	76%	24%	100%
6-12 years	754	397	1,151	66%	34%	100%
13-17 years	158	390	548	29%	71%	100%
Aboriginal Status						
Non-Aboriginal	1,893	842	2,735	69%	31%	100%
Aboriginal	924	387	1,311	70%	30%	100%

Characteristics	Eligible	Not Eligible	Total	Eligible	Not Eligible	Total
Placement Setting						
Foster care	1,411	490	1,901	74%	26%	100%
Relative/Kinship: non-Aboriginal	1,041	384	1,425	73%	27%	100%
Relative/Kinship: Aboriginal	274	165	439	62%	38%	100%
Residential care	47	39	86	55%	45%	100%
Other	44	151	195	23%	77%	100%
ROSH reports						
0 report	99	172	271	37%	63%	100%
1-2 reports	623	262	885	70%	30%	100%
3-6 reports	814	299	1,113	73%	27%	100%
7-15 reports	812	373	1,185	69%	31%	100%
Over 16 reports	469	123	592	79%	21%	100%
Substantiated Maltreatment Reports						
0	533	483	1,016	52%	48%	100%
1	1,307	438	1,745	75%	25%	100%
2	537	139	676	79%	21%	100%
3	247	70	317	78%	22%	100%
4	109	26	135	81%	19%	100%
5	35	9	44	80%	20%	100%
6	20	6	26	77%	23%	100%
7	5	0	5	100%	0%	100%
8	1	0	1	100%	0%	100%
9	1	0	1	100%	0%	100%
Missing	22	58	80	28%	73%	100%
Predominant Maltreatment Issues						
Mixed maltreatment issues	909	469	1,378	66%	34%	100%
Emotional abuse	122	70	192	64%	36%	100%
Neglect	819	260	1,079	76%	24%	100%
Physical abuse	562	174	736	76%	24%	100%
Sexual abuse	65	37	102	64%	36%	100%
No maltreatment issue	318	161	479	66%	34%	100%
Missing	22	58	80	28%	73%	100%

Characteristics	Eligible	Not Eligible	Total	Eligible	Not Eligible	Total
District Offices						
Central Coast	135	44	179	75%	25%	100%
Far West	33	17	50	66%	34%	100%
Hunter New England	531	243	774	69%	31%	100%
Illawarra Shoalhaven	196	69	265	74%	26%	100%
Mid North Coast	144	66	210	69%	31%	100%
Murrumbidgee	153	78	231	66%	34%	100%
Nepean Blue Mountains	217	73	290	75%	25%	100%
Northern New South Wales	142	94	236	60%	40%	100%
Northern Sydney	39	38	77	51%	49%	100%
South Eastern Sydney	148	71	219	68%	32%	100%
South Western Sydney	384	132	516	74%	26%	100%
Southern New South Wales	56	45	101	55%	45%	100%
Sydney	120	52	172	70%	30%	100%
Western New South Wales	242	124	366	66%	34%	100%
Western Sydney	277	83	360	77%	23%	100%

4.3 Completed Interviews

In this section, we examine interview status relative to final orders/study eligibility. Caregivers of children on final orders were recruited to participate in the POCLS. In some cases, caregivers of children on final orders agreed to be interviewed; in other cases, caregivers elected not to participate. Among those caregivers who agreed to have their details passed to the data collection agency, some ultimately declined an interview. The interviewed cohort consists of those children whose caregivers participated in the interview at Wave 1. Overall, as a fraction of the final orders/study eligible cohort, interviews were completed with about 45% of the children with a final care and protection order.

As with final orders/study eligibility, interview status (interviewed/not interviewed) varied with characteristics of the children (see Table 3.2). Notably, interviews were strongly correlated with length of stay (restoration cases were excluded from an interview at Wave 1). Caregivers with children/young people who had been in care for more than two years were much more likely to complete the interview (above 50%). Among children in care for less than two years, completion rates were below 30% .

Gender was not a determining factor in whether an interview was completed. Age, however, was strongly correlated with completion. More than half of the children who entered care as infants had a completed interview; among 13 to 17 year olds, only 28% completed the interview.

Aboriginal status did not influence interview rates but placement setting was important. Foster carers and non-Aboriginal kinship providers were among the most likely to complete an interview. Children and young people in other settings were less likely to complete the interview.

Maltreatment history presents a complicated narrative. Except for children with no ROSH reports, the number of ROSH reports was not correlated with interview completion. With respect to the number of substantiated reports, no clear picture emerged in the data. For predominant maltreatment issues reported, children who experienced sexual abuse are less likely to be interviewed.

District Office was also associated with completion rates. Among District Offices with more than 50 children and young people on final orders, completion rates varied from 33% to 56%.

Table 3.2: Pathways of Care Longitudinal Study Sample by Interview Status and Selected Child Characteristics

Characteristics	Interviewed	Not Interviewed	Total	Interviewed	Not Interviewed	Total
Total	1,282	1,535	2,817	46%	54%	100%
Time in Care						
Less than 1 month	37	103	140	26%	74%	100%
1 to 2 months	4	45	49	8%	92%	100%
2 to 3 months	2	28	30	7%	93%	100%
4 to 6 months	13	107	120	11%	89%	100%
7 to 12 months	26	253	279	9%	91%	100%
13 to 24 months	56	202	258	22%	78%	100%
Over 24 months	1,144	797	1,941	59%	41%	100%
Gender						
Female	642	727	1,369	47%	53%	100%
Male	640	808	1,448	44%	56%	100%
Age at Placement						
Infants	473	408	881	54%	46%	100%
1-5 years	477	547	1,024	47%	53%	100%
6-12 years	288	466	754	38%	62%	100%
13-17 years	44	114	158	28%	72%	100%
Aboriginal Status						
Non-Aboriginal	838	1,055	1,893	44%	56%	100%
Aboriginal	444	480	924	48%	52%	100%

Characteristics	Interviewed	Not Interviewed	Total	Interviewed	Not Interviewed	Total
Placement Setting						
Foster care	653	758	1,411	46%	54%	100%
Relative/Kinship: non-Aboriginal	482	559	1,041	46%	54%	100%
Relative/Kinship: Aboriginal	117	157	274	43%	57%	100%
Residential care	17	30	47	36%	64%	100%
Other	13	31	44	30%	70%	100%
ROSH Reports						
0 report	25	74	99	25%	75%	100%
1-2 reports	286	337	623	46%	54%	100%
3-6 reports	401	413	814	49%	51%	100%
7-15 reports	357	455	812	44%	56%	100%
Over 16 reports	213	256	469	45%	55%	100%
Substantiated Maltreatment Reports						
0	211	322	533	40%	60%	100%
1	608	699	1,307	47%	53%	100%
2	276	261	537	51%	49%	100%
3	96	151	247	39%	61%	100%
4	51	58	109	47%	53%	100%
5	15	20	35	43%	57%	100%
6	14	6	20	70%	30%	100%
7	5	0	5	100%	0%	100%
8	1	0	1	100%	0%	100%
9	0	1	1	0%	100%	100%
Missing	5	17	22	23%	77%	100%
Predominant Maltreatment Issues						
Mixed maltreatment issues	385	524	909	42%	58%	100%
Emotional abuse	58	64	122	48%	52%	100%
Neglect	406	413	819	50%	50%	100%
Physical abuse	262	300	562	47%	53%	100%
Sexual abuse	23	42	65	35%	65%	100%
No maltreatment issue	143	175	318	45%	55%	100%
Missing	5	17	22	23%	77%	100%

Characteristics	Interviewed	Not Interviewed	Total	Interviewed	Not Interviewed	Total
District Offices						
Central Coast	68	67	135	50%	50%	100%
Far West	8	25	33	24%	76%	100%
Hunter New England	273	258	531	51%	49%	100%
Illawarra Shoalhaven	79	117	196	40%	60%	100%
Mid North Coast	66	78	144	46%	54%	100%
Murrumbidgee	83	70	153	54%	46%	100%
Nepean Blue Mountains	82	135	217	38%	62%	100%
Northern New South Wales	75	67	142	53%	47%	100%
Northern Sydney	10	29	39	26%	74%	100%
South Eastern Sydney	63	85	148	43%	57%	100%
South Western Sydney	143	241	384	37%	63%	100%
Southern New South Wales	30	26	56	54%	46%	100%
Statewide Services	40	80	120	33%	67%	100%
Sydney	136	106	242	56%	44%	100%
Western New South Wales	126	151	277	45%	55%	100%

4.4 Multivariate Models

Tables 3.3 and 3.4 show the results of multilevel models[4] for eligibility/final orders and interview completion. Although consistent with what has already been reported, the models clarify important relationships.

With respect to eligibility/final orders, gender was not important, as already noted. Age, however, was an important factor, even after accounting for other child characteristics. Relative to children of other ages, infants were much more likely to receive final orders. Aboriginal status did not influence the likelihood that a child would receive final orders. However, children placed in kinship homes, both Aboriginal and non-Aboriginal, were less likely to become study eligible when compared to children in non-kin foster homes.

[4] The levels involved in the model are child and district office.

Table 3.3: Coefficients of Multilevel Logit Models of Study Eligibility

Variable Name	Log Odds	Standard Error	Probability Value	Odds Ratio
Intercept	3.7794	0.264	<.0001	
Gender				
Female	-0.1037	0.079	0.1912	0.902
Male	Reference			
Age				
1-5 years	-1.3534	0.134	<.0001	0.258
6-12 years	-3.3007	0.169	<.0001	0.037
13-17 years	-1.9937	0.139	<.0001	0.136
Infants	Reference			
Aboriginal Status				
Non-Aboriginal	0.0852	0.101	0.4011	1.089
Aboriginal	Reference			
Placement Setting				
Residential care	0.01688	0.090	0.8518	1.017
Relative/Kinship: non-Aboriginal	-0.4767	0.144	0.0009	0.621
Relative/Kinship: Aboriginal	-1.2814	0.222	<.0001	0.278
Other	0.3102	0.265	0.2413	1.364
Foster care	Reference			
ROSH Reports				
0 report	-2.932	0.256	<.0001	0.053
1-2 reports	-1.7025	0.158	<.0001	0.182
3-6 reports	-1.0599	0.141	<.0001	0.346
7-15 reports	-0.8977	0.132	<.0001	0.407
Over 16 reports	Reference			
Predominant Maltreatment Issues				
Mixed maltreatment issues	-0.3367	0.193	0.0803	0.714
Emotional abuse	-0.4255	0.241	0.0768	0.653
Neglect	-0.00691	0.195	0.9718	0.993
Physical abuse	-0.05541	0.195	0.7766	0.946
Sexual abuse	-0.04304	0.292	0.8826	0.958
No maltreatment issue	Reference			

The number of ROSH reports was associated with receiving final orders/study eligibility, with children having 16 or more ROSH reports being much more likely to receive final orders. Moreover, as the number of ROSH reports increases, the likelihood of reach eligibility increases linearly in that, although the likelihood is always lower, as the number of ROSH reports grows, the differences with children who 16 or more ROSH reports become smaller. Children with a predominant maltreatment issue of emotional abuse or mixed maltreatment issues were somewhat less likely to be eligible than children with no maltreatment issues listed. This is counter-intuitive and may represent underlying issues with the data captured in the KiDS system.

Multilevel results for interview status are reported in Table 3.4. These data show that the gender of foster children is not a statistically significant determinant of whether a child was interviewed, given their eligibility. Age was a significant factor. Carers of older children, especially teenagers, were less likely to be interviewed when compared to infants.

Table 3.4: Coefficients of Multilevel Logit Models of Eligible Children Who Are Interviewed vs. Not Interviewed

Category	Log Odds	Standard Error	Probability Value	Odds Ratio
Intercept	0.464	0.225	0.040	
Gender				
Female	0.107	0.078	0.170	1.113
Male	Reference			
Age				
1-5 years	-0.490	0.112	<.0001	0.613
6-12 years	-1.340	0.224	<.0001	0.262
13-17 years	-0.879	0.128	<.0001	0.415
Infants	Reference			
Aboriginal Status				
Non-Aboriginal	-0.074	0.096	0.440	0.929
Aboriginal	Reference			
Placement Setting				
Residential care	0.362	0.346	0.295	1.436
Relative/Kinship: non-Aboriginal	0.100	0.086	0.245	1.105
Relative/Kinship: Aboriginal	-0.236	0.153	0.123	0.790
Other	-0.372	0.356	0.297	0.689
Foster care	Reference			
ROSH Reports				
0 report	-1.183	0.296	<.0001	0.306
1-2 reports	-0.436	0.155	0.005	0.647
3-6 reports	-0.132	0.133	0.318	0.876
7-15 reports	-0.171	0.123	0.164	0.842
Over 16 reports	Reference			
Predominant Maltreatment Issues				
Mixed maltreatment issues	-0.108	0.168	0.520	0.897
Emotional abuse	0.072	0.235	0.759	1.075
Neglect	0.145	0.166	0.381	1.156
Physical abuse	-0.088	0.163	0.589	0.916
Sexual abuse	-0.281	0.306	0.358	0.755
No maltreatment issue	Reference			

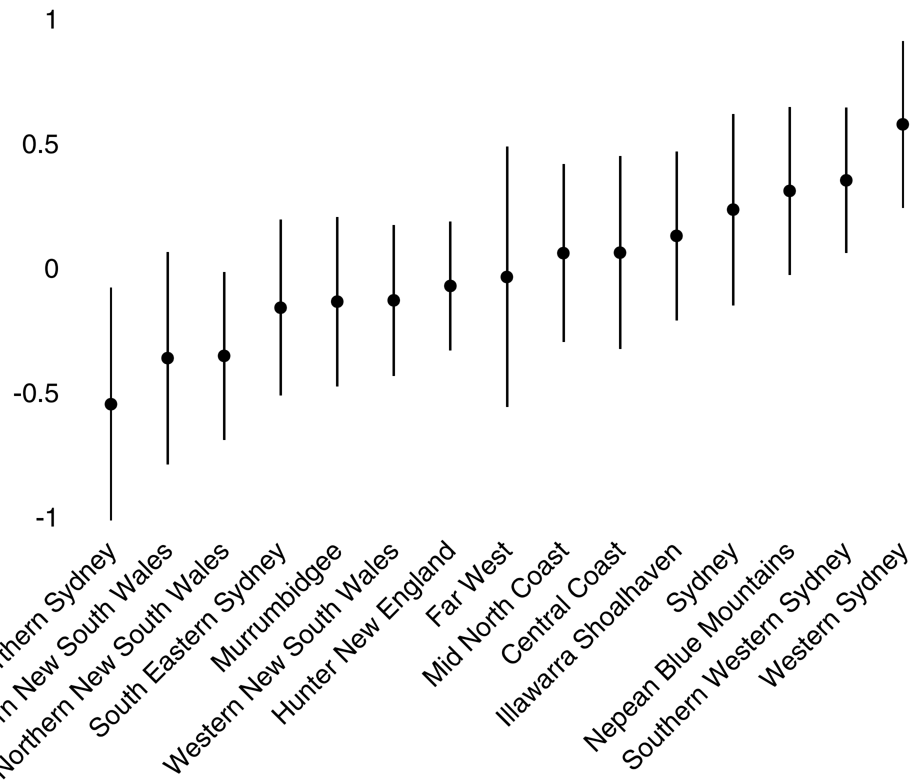
A child's Aboriginal status was not significantly related to the probability of being interviewed. Neither was the placement setting. With regard to ROSH reports and maltreatment type, fewer ROSH reports were associated with a lower probability of having a completed interview. However, maltreatment type was not a factor that changed the likelihood of being interviewed.

4.5 District Office Differences

The descriptive results suggested that District Offices differed with regard to both final orders/study eligibility and interview completion. To test whether these differences were statistically meaningful, we compared the expected final orders/eligibility and interview rates with the observed rates, given the characteristics of children served in each district. The results, which are measured as model residuals, are found in Figures 2 and 3.

Study eligibility at the District Office level is displayed in Figure 2. Depicted as a point estimate is the departure of the observed rate from the expected rate (i.e., the residual) for each District. The vertical line through each point estimate is the confidence interval, which expresses the extent to which the departure is statistically meaningful. In cases where the vertical line crosses the y-axis at zero, the expected and the observed final orders/eligibility rates are not meaningfully (i.e., statistically) different.

Figure 2: Observed and Expected Study Eligibility Rates by District Office



The results indicate that Northern Sydney and Northern NSW had a lower than expected final orders/study eligibility rate whereas South Western Sydney and Western Sydney had above average final orders/eligibility rates. Because final orders/eligibility and length of stay are correlated, the higher rates of eligibility in these areas may be indicative of generally longer lengths of stay in these areas.

Figure 3: Observed and Expected Interview Participation Rates by District Office

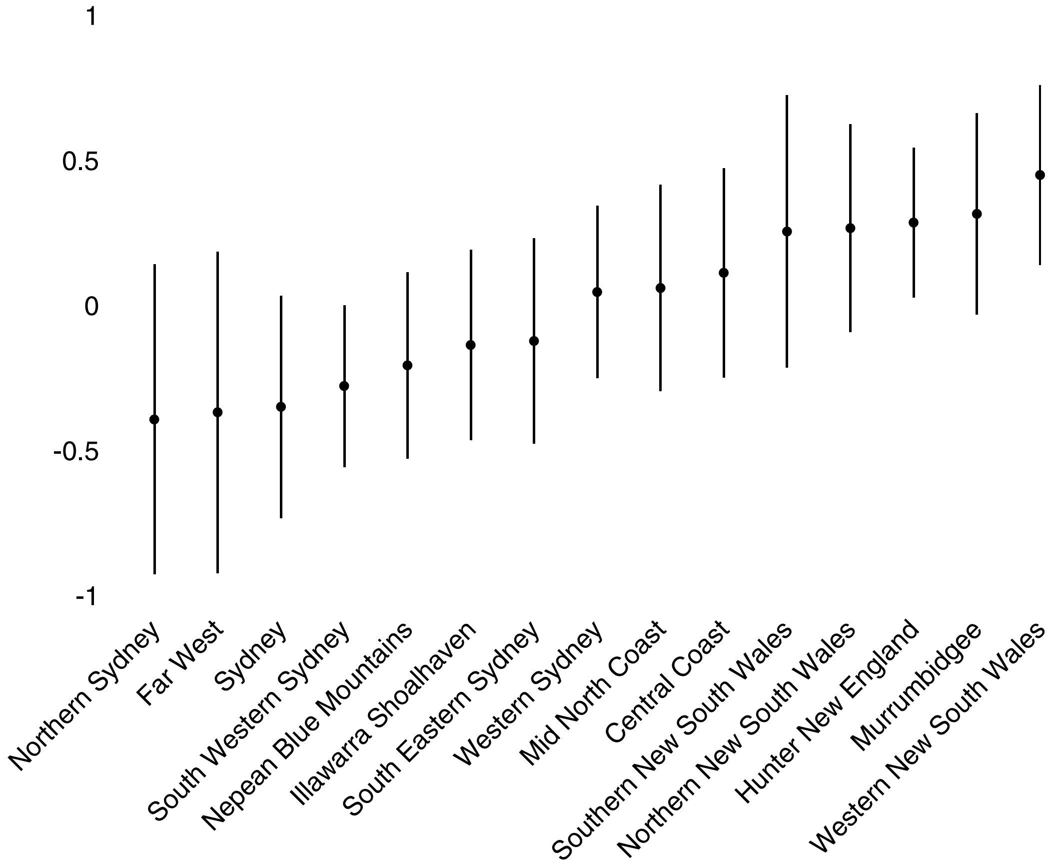
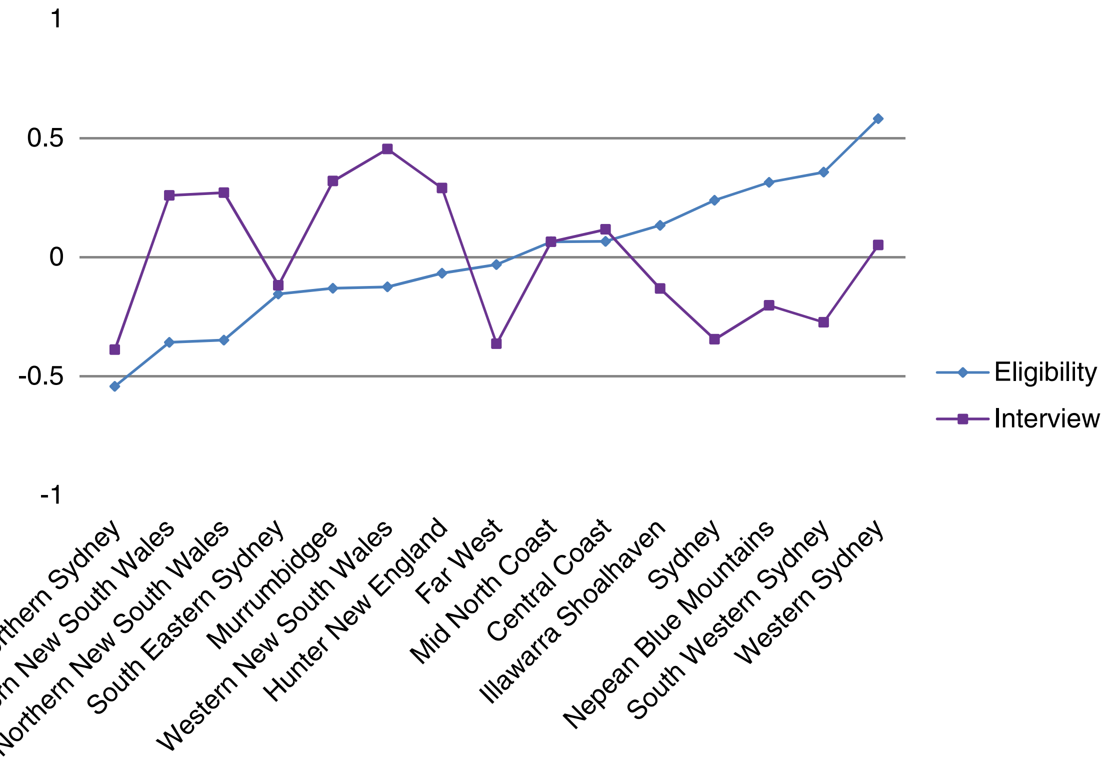


Figure 3 shows the same data for interview rates. These data indicate that none of the District Offices has significantly lower survey participation rates, although South Western Sydney and Sydney are borderline cases. However, the districts of Hunter New England and Western NSW had better than expected participation rates.

Finally, Figure 4, which combines the data in Figures 2 and 3, shows the relationship between final orders/eligibility and participation rates. Northern Sydney had both low final orders/eligibility and participation rates. Western Sydney had an elevated final orders/eligibility rate and an average participation rate. Western NSW had an average eligibility rate and an above average participation rate. Generally, the correlation between eligibility and whether a young person participated in the interview was negative, suggesting that District Offices with higher eligibility rates had lower interview rates.

Figure 4: Study Eligibility and Interview Participation Rates by District Office



4.6 Summary

The POCLS focuses on the developmental wellbeing of children placed on final care and protection orders. As children placed on final care and protection orders are a subset of all children who enter care, it is important to understand who, among all the children who enter out-of-home care, reaches the point of having a final care and protection order before study findings are generalised to subsequent cohorts of children.

Generally, the findings suggest that children with more contact with the child protection system, as measured by the number of ROSH reports, were both more likely to receive final orders and participate in the interviews. The connection between contact and participation is more or less expected. Children on final orders likely come from situations wherein the likelihood of restoration is low. Hence the need for a long-term care and protection order. Children in these situations tend to stay in care longer and the underlying challenges that parents face may be reflected in the fact that prior to entry into out-of-home care they had more ROSH reports and more substantiated prior reports.

We can as a result expect to find that children in these circumstances will have developmental outcomes at Wave 1 that are generally lower than what one might find in the population of out-of-home care children who enter care and then leave placement quickly. This is likely the case with older children who were living at home for longer periods prior to coming into care. The results shown here point to how one might improve the care available to children across the range of placement experiences but especially for children on long-term care orders.

Although the findings presented here do not address the issue directly, the District Office differences in eligibility/final order rates are likely the result of differences in length of stay. Because the model results account for child-level differences (e.g., age and gender) and differences in the child protection history (e.g., number of ROSH reports and reported maltreatment issues), differences in length of stay may be due to differences in restoration rates. A firm conclusion regarding length of stay and restoration requires a more focused study of the study eligible population relative to the larger population cohort, but the benefits of such a study from a policy and practice perspective are likely significant. Typically, when similar children have different placement experiences, the reasons why may have to do with between District Office differences in the approach to care and protection. Evidence of this sort, linked to a rigorous continuous quality improvement process, may improve outcomes.

Finally, the findings also raise a question about sample weights. The results here suggest that generalisations from the survey respondents to the larger study eligible cohort may require the use of sample weights to adjust for non-response bias. This question is taken up in the final section of this technical report.

5 Adjusting for Non-Response Bias in the POCLS Sample

5.1 Introduction

The survey used for the POCLS has missing data. Encountering missing data is not a unique issue for the POCLS; missing data is a recurring issue in survey research. There are generally two different types of missing data in surveys: unit non-response and item non-response. Unit non-response refers to situations wherein study eligible members of a population do not participate. Item non-response refers to questions that were not answered by respondents. Unit non-response results in a survey population with two potentially different sub-populations: respondents and non-respondents.

Unit non-response complicates statistical analysis because information is completely missing for non-respondents. Generally, the collected survey data provides information on surveyed sample units, not the population. If conclusions drawn from the sample apply only to the sample, statistical adjustments are not needed. However, issues arise when conclusions are reached and then generalised to the population as a whole. In the case of non-response, if unit non-response were to occur randomly, this would be a special case in which survey units have an equal chance to be selected. In this case, it is possible to generalise from sample statistics to the whole population. However, survey results may not be generalisable to the whole population if unit non-response is non-random. Unit non-response may result in over- or under-represented groups because the respondents differ from non-respondents regarding observables (i.e., demographic characteristics) and/or unobservables (i.e., motivation) that are correlated with survey outcomes, leading to estimates that are in some way biased. The issue can escalate when response rates are relatively low[5]. The challenge is to determine how to approximate a representative sample using some statistical adjustments given non-response bias.

In what follows, we discuss the issue of non-response bias in the POCLS and propose ways to think about how the POCLS data can be used. The discussion specifically covers: POCLS survey non-response, assumptions on unit non-response, non-response bias analysis results, and two types of weights (cell weights and propensity weights) for balancing the POCLS sample.

[5]National Center for Education Statistics (2016) claims that ‘any survey stage of data collection with a unit or item response rate less than 85% must be evaluated for the potential magnitude of non-response bias before the data or any analysis using the data may be released’.

5.2 Data

The extent of non-response in the POCLS survey is well documented (Paxman et al., 2014). Their report describes the POCLS survey non-response in terms of the sample frame. According to Paxman et al., there are three hierarchical groups of people. At the time we worked with the data these were: the population cohort, the study eligible cohort, and the survey cohort.

- Population Cohort: The total population cohort (N=4,126) consisted of all children aged 0-17 entering into out-of-home-care for the first time.
- Study Eligible Cohort: Out of the total population, 2,827 children were included in the study eligible cohort. Eligibility was based on whether a final care and protection order granting custody to the Minister had been filed with the Court.
- Survey Cohort: The survey cohort was smaller than the study population because caregivers had to consent to the survey. A subset of children (N=1,788) agreed to participate in the survey. In this sample, some of children returned to their parents and were not invited to participate in the survey for pragmatic and ethical reasons. Eventually, 1,285 children (and their carers) participated in the Wave 1 survey.

For purposes of this analysis, children without an identified District Office or identified with 'statewide services' were excluded. Also, children whose placement was respite care were also excluded. With these exclusions, we determined that the total number of children was 4,068; the study eligible cohort consisted of 2,826 in out-of-home care. Within that group, 1,787 children agreed to participate in the survey and 1,284 children (45.4% = 1,284/2,826) actually responded to the survey. The survey respondents represent the sample drawn for purposes of generalising to the study eligible cohort.

5.3 Assumptions on Unit Non-response

Before discussing the statistical techniques used to address non-response bias, it is important to lay out statistical assumptions associated with non-response. When addressing unit non-response, the following three distinctive assumptions on unit missing data defined by Little and Rubin, 2002 provide a general framework for assessing the current issue and guiding potential solutions to overcome non-response bias:

- Missing Completely at Random: If the probability of unit response is independent of the survey outcomes, this is a situation called missing completely at random (MCAR). This is the most favourable situation and there is no need to adjust for unit non-response. However, this situation is highly unlikely in most surveys.
- Missing at Random: If the probability of unit response is independent of the survey outcomes after accounting for auxiliary variables, this is a

situation called missing at random (MAR). MAR is frequently assumed in other surveys and is the assumption for most weighting adjustments.

- Missing Not at Random: If the probability of unit response is still dependent on survey outcomes even after accounting for auxiliary variables, this is called missing not at random (MNAR). MNAR is the most complicated situation because the dependency cannot be adjusted for by any auxiliary variables.

5.4 Non-response Bias Analysis

Unless we assume MCAR, it is hard to apply with confidence the conclusions drawn from the sample to the target population due to the initial differences between the two groups (respondents and non-respondents) in terms of survey outcomes. Statistical adjustments to the sample (i.e., weighting) raises the confidence.

We performed a non-response bias analysis to determine whether the MCAR condition is satisfied. Ideally, survey outcomes would be used to check unit non-response bias; however, those outcomes are not available because they did not respond. This is the same reason that a non-response bias analysis using auxiliary variables was used. Even though survey outcomes are not available for non-respondents, administrative data (i.e., gender, age, race/ethnicity, etc.) are available for both respondents and non-respondents.

These data open the door to comparisons between respondents and non-respondents using auxiliary variables. We included the following auxiliary variables for non-response bias analysis as follows:

- Type of the first placement, which was categorised into three: foster care, kinship care, and others (group care and other care types not listed separately).
- Age at the start of the placement, which was categorised into three groups: age 0, age 1 through 5, and above age 5.
- Gender.
- Aboriginal status. Both Aboriginal and Torres Strait Islander are combined and treated as Indigenous.
- Maltreatment issues. The predominant type of primary and secondary reported maltreatment issues of CP episodes initiated prior to the spell was included. (1) Issue-None indicates no maltreatment issue, (2) Issue-Neglect indicates more than 50% of the reported issues are neglect, (3) Issue-Abuse indicates more than 50% of the reported issues are physical abuse, sexual abuse, or emotional abuse, and (4) Issue-Mixed indicates mixed maltreatment issues.
- The number of ROSH reports, initiated prior to admission into out-of-home care, ranged from 0 to 49. Approximately 70% were from 0 to 10. Five categories were created using the continuous data: (1) indicates missing

or zero ROSH reports, (2) indicates 1 or 2 ROSH reports, (3) indicates 3 through 6 ROSH reports, (4) indicates 7 through 15 ROSH reports, and (5) indicates 16 or more ROSH reports.

Table 4.1 shows the results of non-response bias results. Considering the average response rate of 45.4%, over- or under-representations can be found by comparing the response rates of individual categories with the average rate (45.4%).

Table 4.1: Response Rates by Auxiliary Variables

Auxiliary Variable	N	Percent Responding
Average	1,284	45.4%
Care Type		
Foster family care	1,815	45.0%
Kinship care	719	45.6%
Other care types	292	47.9%
Age at Admission		
Under 1	883	53.7%
1 to 5	1,026	46.5%
6 and above	917	36.3%
Indigenous Status		
Non-Indigenous	1,900	43.9%
Indigenous	926	48.6%
Gender		
Female	1,375	47.1%
Male	1,451	43.9%
Maltreatment Type		
No ssue specified	398	47.0%
Neglect only	844	50.1%
Abuse only	709	41.7%
Mixed maltreatment types	875	43.2%
Number of ROSH Reports		
No ROSH reports	98	22.4%
1 to 2 ROSH reports	620	45.8%
3 to 6 ROSH reports	819	49.5%
7 to 15 ROSH reports	817	43.9%
More than 16 ROSH reports	472	45.3%

Regarding care type, response rates did not vary across different care types with the exception of a slightly higher response rate in 'Type-others' (47.9%). Respondents are much younger than non-respondents because more under 1 year olds responded (53.7%) and fewer older children (6 and above) responded (36.3%). More Indigenous children (48.6%vs. 43.9%) and females (47.1%vs. 42.9%) responded. Based on the results, respondents are more neglected and non-respondents are more abused. More respondents are in the 3 through 6 ROSH report categories.

We conducted a more formal non-response bias test using multivariate modelling that incorporated a response dummy variable as a dependent variable. If the MCAR condition is satisfied, the estimates for covariates are close to zero and odds-ratios are close to one. However, if differences between respondents and non-respondents are found using auxiliary variables, we would conclude that the MCAR condition was not satisfied and therefore non-response bias exists.

Table 4.2 shows the non-response bias test results. Regarding care type, there is no statistically discernible difference among them. Relative to children under the age of 1, fewer children between the ages of 1 and 5 responded; even fewer children age 6 and above responded. After accounting for other variables, no non-response biases were observed in Aboriginal children and male children. In terms of maltreatment issues, abused children were less likely to respond; however, regarding neglect and mixed maltreatment, the results showed a similar response likelihood. Regarding the number of ROSH reports, children who have ROSH report history showed more likelihood to respond. Based on the test results, children who responded were different in terms of children spell age, abuse maltreatment history, and the existence of ROSH reports. Therefore, we concluded that the MCAR assumption does not hold.

Table 4.2: Results of Non-response Bias Test

Auxiliary Variables	Estimate	Standard Error	Pr. > Chi Sq.	Odds-Ratio
Intercept	-0.8454	0.256	0.001	
Care Type				
Foster family care	Baseline			
Kinship care	0.1066	0.0914	0.2436	1.112
Other care types	-0.0264	0.1354	0.8453	0.974
Age at Admission				
Under 1	Baseline			
1 to 5	-0.4918	0.1121	<.0001	0.612
6 and above	-0.9413	0.1235	<.0001	0.39
Indigenous Status				
Non-Indigenous	Baseline			
Indigenous	0.082	0.0827	0.3217	1.085

Auxiliary Variables	Estimate	Standard Error	Pr. > Chi Sq.	Odds-Ratio
Gender				
Female	Baseline			
Male	-0.1213	0.0773	0.1167	0.886
Maltreatment Type				
No issue specified	Baseline			
Neglect only	-0.00913	0.1557	0.9533	0.991
Abuse only	-0.3636	0.1499	0.0153	0.695
Mixed maltreatment types	-0.2212	0.1596	0.1659	0.802
Number of ROSH Reports				
No ROSH reports	Baseline			
1 to 2 ROSH reports	1.0539	0.2735	0.0001	2.869
3 to 6 ROSH reports	1.4094	0.2821	<.0001	4.094
7 to 15 ROSH reports	1.384	0.289	<.0001	3.991
More than 16 ROSH reports	1.606	0.2981	<.0001	4.983

5.5 Non-response Weight Methods

Once it is determined that the MCAR condition is not satisfied, statistical weighting may be considered as a way to adjust for potential bias caused by differential non-response. As shown before, certain types of people are more or less likely to respond to a survey. For example, the survey cohort has more infants than the study eligible population, which might lead to biased estimates. Knowing that both groups (respondents and non-respondents) are systematically different in terms of certain characteristics, use of the auxiliary variables allows us to balance out the over-or under-represented samples.

This approach assumes MAR and thus this indicates that any systematic differences between respondents and non-respondents are unrelated to unobservables after accounting for the observables. Therefore, once any observed differences are adjusted, both respondents and non-respondents are expected to have the same statistical behaviour (in this case, the survey outcomes) because now the sample is representative of the reference population. Once the weights are determined, survey estimates can be obtained using the weighted sample, not the collected sample.

5.6 Probability (Cell) Weights

In order to approximate the target population using the surveyed children, probability (cell) weights were employed first.

Mutually exclusive adjustment cells were created for the probability weights based on auxiliary variables drawn from administrative data.[6] All children within the same cell were given the same weights. Having more variables leads to more cells and more homogeneous groups, which reduces bias even more; however, this can also result in empty cells or cells with extremely few observations. In that case, some cells were collapsed with others.[7]

Table 4.3: The Count of Samples in Two Variables Taking Three Values

		X_1		
X_2		0	1	2
	0	n_1	n_2	n_3
	1	n_4	n_5	n_6
	2	n_7	n_8	n_9

In order to illustrate probability weighting adjustment, Table 4.3 is presented using two variables (X1 and X2). The count for different combinations of X1 and X2 in the sample is denoted as nj (where j=1 to 9) For example, n1 indicates the number of sample units with X1=0 and X2=0. This cross-work can be expanded to more than two variables. If there are three variables that have four, three, and two categories, respectively, then 24 (=4*3*2) mutually exclusive cells can be created. Also, this approach applies to the study population (Nj denotes the number of units in the target population). Once nj and Nj are calculated, for each cell, [1/(nj /Nj)]= (Nj /nj) becomes an adjustment factor to compensate for non-respondents.

After employing probability weights, it is likely that representative estimates can be drawn because both under-represented samples and over-represented samples can be adjusted by putting more weight on under-represented samples and putting less weight on over-represented samples, which essentially means the units in each cell have an equal chance to respond. With weighting, the distribution of surveyed samples can approximate the distribution of the study population. Finally, note that this adjustment is made based on the assumption of MAR. We expect that individuals in each mutually exclusive cell will behave similarly with respect to the survey results.

5.7 Propensity Weights

The alternative non-response adjustment that we employed is propensity weighting, which has gained traction recently as an approach to non-response bias. Propensity weighting uses logit or probit model to estimate response probability or the propensity score. Technically, a response dummy indicator (response or non-response) is regressed on a set of auxiliary variables similar

[6] The auxiliary variables used for weights are the same as the variables in non-response bias analysis.

[7] After collapsing, a total of 122 cells were formed.

to those used in probability weighting.[8] Propensity scores can be defined as the probability of being assigned to the response group depending on the auxiliary variables. The conditional probability can be denoted as follows: $P(R=1|X=x)$, where R is response status (1 if responded, 0 if not responded) and X is a vector of auxiliary variables. Propensity score is an estimate for P. Because respondents elected to participate, their propensity to participate is more in line conceptually with the weighting scheme.

If samples have the same observables, they will have the same propensity score. Conditional on the propensity score, the distributions of auxiliary variables become the same for both respondents and non-respondents (Rubin, 1997). Using propensity weighting, respondents and non-respondents can be equally represented, which is what one needs to draw inferences on the whole population using only respondents. The coefficients obtained from logistic regression show logits (log-odds) as follows: $\text{Logit} = \text{Log (odds)} = \log(p/(1-p)) = X\beta$. Therefore, p (propensity score) can be calculated as follows: $p= 1/(1 + \exp(-X\beta))$. The inverse of the propensity score is used for as the weighting adjustment factor.

5.8 Propensity Weights vs. Probability Weights

Propensity weights have a couple of advantages compared to probability weights. First, they are not limited to cell sizes and ratios. Second, the adjustment factor tends to be more stable and smoother than the probability weights (Carlson & Williams, 2001). In practical terms, the larger probability weight means that respondents in the sample with those characteristics will be weighted to a more significant degree than would be true if the propensity weights were used.

5.9 Normalisation

Even though propensity weights are relatively more stable compared to probability weights, survey units that have extreme propensity scores can still have a substantial influence on survey results because survey outcomes are sensitive to them. To overcome the instabilities caused by extremely high weights of survey units from longer-tailed distributions, weight trimming was one of the options considered. However, determining an optimal cut-point for trimming is not clear cut; trimming also makes weighted sample less representative of the population.

For the POCLS, normalisation was employed to preserve the sample size. The raw propensity weights were normalised by dividing the propensity weight by its overall mean weight (propensity weight/mean weight). Because sampling weights were adjusted by the base, the sum of the raw weights is the population size (2,826). However, normalising the weights makes the

[8] All the interaction terms of auxiliary variables were used regardless of statistical significance.

weight sum to the sample size (1,284), which produces estimates of standard errors and significance tests that are accurate.

In the end, three non-response adjustment weights were calculated for each respondent based on the methods described. The three weights are highly correlated and we expect that the three weights will produce similar results for the POCLS survey. However, due to stability, propensity weights may be a preferred method relative to probability weights. Moreover, in order to preserve the sample size for correct standard error estimates, normalised propensity weights are preferred to non-normalised propensity weights.

5.10 Limitations and Recommendations

Both probability weights and propensity weighting assume MAR. It is also assumed that propensity weighting using a set of observables to calculate the propensity score is sufficient to deal with non-response bias. However, even though samples become representative in terms of observable variables, there is no guarantee that an adjustment using probability weights or propensity weights can adjust for hidden unobservable factors, especially considering the voluntary nature of participating in the Wave 1 survey. It is also important to bear in mind that the benefits of weighting depend on how the auxiliary variables are correlated with the outcomes of interest. With the record linkage, it may be possible to test the correlation between the auxiliary variables used to weight the sample, provided variables in the linkage data sets are correlated with the survey outcomes (i.e., the wellbeing measures that are at the core of the POCLS).

With regard to whether the sample should be weighted, there are two considerations. In the event a user of the POCLS data is interested in making inferences about the sample, then weights are less important. The smaller sample size reduces statistical power, which means one is less likely to find significant relationships. At the same time, it is important to bear in mind that half the children in the study eligible sample are found in the POCLS sample.

As a group, they represent a substantial subset of children being cared for in NSW and the POCLS survey provides considerable information about who those children are, how they are doing, and what might be done to improve their life course outcomes. If the auxiliary variables were used to identify similar groups of children and young people in out-of-home care today, one does gain insight into how care today might be adjusted to promote better outcomes. This is an important type of generalisation that is often overlooked, from a policy/practice perspective. If one wishes to generalise to the study eligible population from the respondents, then analysis with and without the weights has to be undertaken so as to reduce the possibility of Type 1and Type 2errors. When any estimation or analysis is carried out using the weights, the analysts should use the complex survey option in whatever software they use to perform the analysis. For example, in SAS, the survey's design features should be incorporated by employing PROC SURVEYFREQ, PROC SURVEYREG or PROC SURVEYLOGISTIC.

6 References

Carlson, B. L., & Williams, S. (2001). A Comparison of Two Methods to Adjust Weights for Non-Response: Propensity Modeling and Weighting Class Adjustments. Presented at the Proceedings of the Annual Meeting of the American Statistical Association, Princeton, NJ.

Christ, S. L., Biemer, P., & Wiesen, C. (2007). Guidelines for Applying Multilevel Modeling to the NSCAW Data.

Cohen, J. (1992). A Power Primer. Psychological Bulletin, 112, 155–159.

Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability, Vol. 57. New York and London: Chapman and Hall/CRC.

Faraway, J. J. (2006). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models (pp. 1–345). Boca Raton, London, New York: Chapman & Hall/CRC.

Faraway, J. J. (2014). Changes to the Mixed Effects Models Chapters in ELM.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. Behavior Research Methods, 39, 175–191.
Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (pp. 1–408). New York, NY: Wiley and Sons.

National Center for Educational Statistics. (2016). Statistical Standards: Processing and Editing of Data (pp. 1–24). Washington, DC: National Center for Educational Statistics.

Paxman, M., Tully, L., Burke, S., & Watson, J. (2014). Pathways of Care: Longitudinal Study on Children and Young People in Out-of-Home Care in New South Wales. Family Matters, 15–28.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. Journal of the Royal Statistical Society Series B-Statistical Methodology, 60, 23–40.

Rubin, D. B. (1997). Estimating Causal Effects From Large Data Sets Using Propensity Scores. Annals of Internal Medicine, 127, 757–763.