Replicating Multiple Findings on the Predictors of Child Socio-Emotional Behaviour in British-born Children: Evidence From the Millennium Cohort Study

Candidate Number: 1030952

Replication Project

Hilary 2020

Table of Contents

List of Tables

List of Figures

1 | Introduction

Child health and development is a complex issue spanning multiple disciplines including public health and epidemiology, sociology, demography, and economics. Inequalities in child behavioural difficulties are pertinent as they are linked to disparities in a range of later outcomes including labour market ones in adulthood (Lundborg, Nilsson, & Rooth, 2014). There have been a number of explanations as to the causes of child behavioural difficulties, arising from socio-economic, demographic, and psychological factors. However, there is little consensus over what control variables ought to be included in statistical models predicting child behavioural development – even though this can change the coefficients and findings produced quite drastically. This has resulted in a highly eclectic body of literature and a considerable dearth of replications on the subject.

The replication of important findings by multiple independent investigators is fundamental to the accumulation of scientific evidence. Yet, replications tend not to be done as they are often time and labour-intensive while producing relatively little effect (G. Christensen, Freese, & Miguel, 2019). To remedy this, more journals (like Demographic Research[1], PLoS One[2], and The BMJ[3]) are encouraging replicability and by incentivising or requiring authors to be transparent with the data and models used in their papers. As such, there have been an increasing number of replications done in the social and epidemiological sciences.

This paper seeks to add to the growing number of replications by tackling one sub-field of the demographic sciences: family demography. It focuses on child socio-emotional behaviour, defined as a composite metric combining social, emotional, and behavioural problems (Goodman, 1997). To date, there has been no significant replication done focusing on the determinants of child

[1]See https://www.demographic-research.org/volumes/replicable_articles.htm
[2]See https://journals.plos.org/plosone/s/data-availability
[3]See https://authors.bmj.com/policies/data-sharing

socio-emotional development using Millennium Cohort Study data, which has longitudinal data for the cohort of British children born mainly in years 2000/2001. This paper seeks to examine the effect of exploiting researcher degrees of freedom in four separate studies.

A researcher degree of freedom refers to the different reasonable and but often atheoretical choices that researchers need to make when conducting research (Gelman & Loken, 2013). For example, while age is an important variable to control for, whether age groups (categorical) or age in years (continuous) ought to be included in a statistical model is less clear-cut. Yet this can have the effect of changing the final reported coefficient size. Using specification curves in a multiverse analysis (Simonsohn, Simmons, & Nelson, 2015), this study looks to capture the effect of varying such researcher degrees of freedom by examining how the coefficient of the key explanatory variables change when such freedoms are exploited.

The two types of researcher degrees of freedom this paper aims to study are on (1) changing the method of constructing certain variables, and (2) the choice to include or exclude control variables on the coefficient of a given explanatory variable. It seeks to verify and reanalyse four existing studies in reputable social science and medical journals like the BMJ and Social Science & Medicine.

The four studies examine the association between parental discipline practices (Rajyaguru, Moran, Cordero, & Pearson, 2019), ethnicity (Zilanawala, Sacker, Nazroo, & Kelly, 2015), family income (Noonan, Burns, & Violato, 2018), and maternal employment (McMunn, Kelly, Cable, & Bartley, 2012) on British children's socio-emotional development. These were chosen as they use the same original data and have the same outcome variable (albeit measured at different time points), lending themselves well to a cohesive set of replications.

Using the same publicly available dataset, its first objective is to reconstruct the same set of variables and covariates before beginning the analysis. Its

second objective is to rerun the same statistical models to achieve the same set of headline coefficient magnitudes and significances. Its third objective is to then extend these models and reanalyse the coefficients of interest after exploiting the two types of researcher degrees of freedom in each study. It presents this multiverse analysis in the form of a specification curve for each of the coefficients of interest.

The rest of the papers is structured as follows: after the introduction, chapter 2 describes the secondary data used (the UK Millennium Cohort Study) and its data collection method. It also lists some of the ways which the same concept may be constructed in different but reasonable ways using the same data.

Chapter 3 reproduces the four studies in separate sections. In each section, the paper briefly describes each study and explains the variables used before performing the replications and multiverse analyses. Finally, chapters 4 and 5 discuss the findings before concluding.

2 | Method

The paper did not re-run existing code or command lines as none were available online. Attempts were made to contact the original authors for these but were met with limited success. Regardless, the four papers were transparent enough about how they constructed their variables and regression models in such a way that replications were in principle possible.

The data wrangling and analysis for this paper was done with open-sourced programming language R (version 3.6.2). Since the four studies had done their previous analyses in Stata, it was not always possible to exactly copy their method such as when performing multiple imputation or calculating marginal effects.

The R script used for this paper is available online[1].

2.1 Data

The four original studies used the same data from the Millennium Cohort Study[2] (MCS), a study based at the Centre for Longitudinal Studies at University College London. Data was retrieved through the UK Data Service.

The MCS is a longitudinal study of 18,818 children born in the UK between September 2000 and January 2002. It is a multidisciplinary survey developed to capture the effects of childhood development and other outcomes (Hansen, 2014). While there have been seven surveys to date (conducted at 9 months of age, followed by at ages 3, 5, 7, 11, 14, and 17), the four studies have only used data up to and including the fifth survey. There were 13,287 children remaining in the study at the end of the fifth survey (Hansen, 2014). The MCS oversampled from areas with higher levels of poverty and higher proportions of ethnic minority groups (Hansen, 2014). At each wave, caregivers

[1]Please see https://github.com/oxfordmphil1030952/Replication_Project
[2]See https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study

were surveyed via interview and self-completed questionnaire.

This paper was a secondary analysis of the MCS, using de-identified data available in the public domain. As this project falls within the remit of the original parent ethics approval (Hansen, 2014), no additional ethics clearance was required to access and download the anonymised dataset.

2.2 Construction of Sample

Despite investigating the same outcome variable, each of the studies included different control variables. Where the same concept was controlled for, the exact variable used was often constructed differently. The main examples of such cases are listed below:

Total Difficulties Score

The outcome variable used to measure child socio-emotional behaviour was the Total Difficulties Score (TDS) of the Strengths and Difficulties Questionnaire[3] (SDQ) (Goodman, 1997). It is a short parent-reported screening questionnaire for psychopathology in the context of emotional and behavioural difficulties (Goodman, 2001). It comprises 25 items (see Appendix A), with five items each covering emotional, conduct, hyperactivity, peer, and prosocial problems. Items from the first four categories were summed to create the TDS, ranging from 0 to 40 points.

There were two ways the studies have used this variable. While some studies used the TDS as a continuous variable with higher scores indicating greater psychopathology, others generated a binary variable with a score of 17 and above (out of 40) indicating problematic socio-emotional behaviour and below 17 indicating normal behaviour (Noonan et al., 2018). The nature of the outcome variable (continuous or binary) has the effect of changing the appropriate statistical model (OLS or logistic regression) used. The interpretation of the coefficients would also change accordingly (marginal effects

[3]See https://www.sdqinfo.com/

or odds ratios) and are thus not directly comparable to one another.

Demographic Variables of Mother and Child

The main demographic variables controlled for throughout the papers were the age and ethnicity of the mother and child.

In the MCS data, there were three different ethnic typologies available for researchers to choose from for both mother and child: a 6, 8, and 11-category typology. Another possible researcher degree of freedom seen was to place 'Mixed' ethnicity respondents into the 'Other' category. Different studies chose different ways of classifying ethnicity, seemingly without a theoretical justification.

Throughout the different waves, there were broadly also two ways of measuring the age of the mother and child. For mothers, age in years was available as a continuous variable, but so too were banded age groups as a categorical variable. In early waves, child age was available in number of days. Child age in years was available in later waves.

Socio-economic Variables of Mother

Two main socio-economic variables used were educational attainment and household/personal income.

The questions for educational attainment were not always consistent. Some waves had plainly asked for the mother's highest educational qualification while others used the National Vocational Qualification (NVQ) equivalence scale with resulting differences in the coding of responses.

The variable for income had much more variation. The researcher degrees of freedom was choosing between using the logarithm of income as a continuous variable or using banded income groups as a categorical variable. For studies using mean income across waves (persistent income), taking the arithmetic mean first and then logarithm second or taking logarithm first and then the arithmetic mean second produced different figures.

Psychological Variables of Child

The variable on infant temperament measured at nine months of age had two ways of being constructed. While the data provided could be used as a continuous variable, it was also possible to construct it as a binary variable indicating unusually high temperament or not.

Mother-reported Social Status

Social status was classified based on the occupation of the mother using the National Statistics Socio-economic Classification (NS-SEC) scale, with a 5, 7, and 13-category typology available in the original data.

Dealing with Missing Variables

All but one of the studies (looking at the association of ethnicity on SDQ TDS by Zilanawala et al.) have used subsetting to remove missing values. Their study has instead used multiple imputation to fill in missing values. This researcher degree of freedom was explicitly modelled, with coefficients in a subsetted and imputed dataset included in the same specification curve for that study.

3 | Replicating the Four Studies

The four studies replicated each examine a different explanatory variable of child socio-emotional behaviour at different ages. This section looks at how the coefficients of the key explanatory variable changes when the researcher degrees of freedom, in constructing and including or excluding control variables, are exploited.

3.1 Study 1: Rajyaguru et al., 2019

Published in the Journal of the American Academy of Child and Adolescent Psychiatry, the study by Rajyaguru et al. examined the association between different parental discipline methods on their child's socio-emotional behaviour. It tested whether a longitudinal association between different disciplinary parenting practices at 3 years of age and later child psychopathology at 11 years existed. 'Active' (smacking, shouting, and telling off) and 'Withdrawal' (ignoring, removal of privileges, and sending to bedroom) approaches of child discipline were distinguished as two separate variables whose coefficients were each studied (Rajyaguru et al., 2019).

3.1.1 Variables and Covariates Used

There were 10 control variables in addition to the 2 key explanatory variables used representing the two main approaches to discipline – Active and Withdrawal discipline methods. Categorical variables were coded and indexed beginning with 0.

Key Predictor Variables

Active or Positive discipline methods are characterised by a more sensitive approach based on warmth (A. Christensen & Heavey, 1987). ***Withdrawal*** or Negative approaches on the other hand are to do with harsh or punitive measures incorporating elements of hostility or psychological con-

trol (Rajyaguru et al., 2019). The MCS wave 2 dataset contained seven items pertaining to disciplinary practice originating from the Conflict Tactics Scale developed to explore inter-familial conflict and parent interaction with children (Straus, 2017). Mothers were asked about their child's behaviour over the past 6 months with the questions: 'How often do you ignore/smack/shout/send to bedroom or naughty chair/take away treats/tell off/bribe with sweets or other ... when [child] is naughty.' Frequencies were coded on a 6-point scale – categories: can't say, never, rarely, once a month, at least once a week, and daily – and re-assigned values 0 to 5 respectively. 'Can't say' was assigned 0 in this paper because while there is no information on frequency given, it cannot be classified as missing data. The original study was not clear on how it treated 'can't say' responses. The respective items were summed into Active (smacking, shouting, and telling off) and Withdrawal (ignoring, removal of treats, and sending to bedroom) continuous variables. Bribing was not assigned to either category, as in the original study.

Outcome Variable

The outcome variable ***SDQ TDS*** at MCS wave 5 was available in the original MCS data and the continuous score was used in its raw form without any re-coding apart from for missing values.

The study also reported including a child-reported mood measure at wave 5 as part of the outcomes it was studying. It was not clear how the study included this variable together with the SDQ TDS, so this paper has omitted this variable and only focused on the TDS as the key outcome variable for consistency with the other three original studies analysed.

Control Variables

An ***Overall Discipline Use*** continuous variable was created to take into account the total 'amount' of discipline used as distinct from the 'type' of discipline used. This variable was constructed from the same raw data used

in the key predictor variables by summing up all seven discipline methods. The study did not provide exact details of its construction. For this paper the scores of all discipline questions were similarly re-coded 0 to 5 and then summed.

Infant Temperament measured at 9 months of age (MCS wave 1) was included as the original study's way of trying to account for reverse causality. There were 14 items designed to capture the temperament of children across four categories: regularity, approach-withdrawal, adaptability, and mood. The original study stated that it summed the individual items to create a continuous variable. The six possible responses were the same for the discipline methods – categories: can't say, never, rarely, once a month, at least once a week, and daily. Similarly, 'can't say' was originally coded as 6 but was re-coded to a value of 0 as above. The 14 individual scores were then summed to create a continuous variable. The original study was again not clear on how it treated 'can't say' responses.

The study used a continuous variable for infant temperament, but like other psychological and psychometric variables, a binary variable could also be created with 0 indicating 'normal' level of temperament and 1 indicating 'abnormal' levels of temperament. Since there was no immediately obvious cutoff, this study used the arithmetic mean of the continuous variable (greater than or equal to 45 coded into 1) as the cutoff to create a binary variable.

There were two variables representing ***Mother's Age*** at childbirth in the original data, one for age in years as a continuous variable and another for banded age. The four groups were 12 to 19, 20 to 29, 30 to 39, and 40 and above. The original study used age as a categorical variable.

Child Sex or gender, a binary variable, was also used as a control variable.

Maternal Depression at MCS wave 5 was also used as a binary control variable, with the respondent being asked if they were ever medically diagnosed with depression or anxiety and the yes/no responses coded.

Maternal Psychosocial Distress at MCS wave 1 was assessed using a modified version of the Rutter Malaise Distress Inventory designed to identify emotional disturbance and associated physical symptoms. 9 items were included in the questionnaire – asking if the respondent often felt tired, miserable or depressed, worried, often gets into violent rages, gets scared for no reason, is easily upset or irritated, if every little thing gets on their nerves, or whose heart often races like mad (phrasing lifted from original questionnaire) – each with a yes/no answer. These yes (coded 1) and no (coded 0) responses were summed to create a continuous score with a possible total of 9 points. Alternatively, it was also possible to create a binary variable with a cut-off score of 4 or more indicating high levels of distress. The original study used the binary variable.

Maternal Self-esteem at wave 1 was similarly derived from a shortened six-item version of the Rosenberg Self Esteem Inventory. It asked respondents if they were satisfied with themselves, was able to do things as well as others, has a positive attitude to oneself, thinks of themselves as no good, feels useless, and feels like a failure. Each question had four responses: Strongly Agree, Agree, Disagree, and Strongly Disagree which were coded or reverse-coded (depending on whether the question was positive or negative) 0 to 3 and summed to construct a continuous variable. Similarly, it was also possible to create a binary variable with a cut-off score of below 9 indicating low levels of self-esteem. The original study used the binary construction of this variable.

Parity refers to whether the mother had other children apart from the child in the MCS birth cohort. Asked in wave 5, this was a binary variable where 1 meant the child had natural siblings while 0 meant they did not.

Maternal Education was coded differently in wave 1 compared to later waves. Wave 1 had eight categories re-coded 0 to 7: higher degree, first degree, diplomas in higher education, A/AS/S levels, O level/GCSE grades A to C, GCSE grades D to G, other qualification including overseas, and

none of these qualifications respectively. Later waves on the other hand had seven categories re-coded 0 to 6: NVQ levels 1 to 5, followed by overseas qualifications only, then none of these respectively. The study used data from wave 1.

Social Status was derived as a function of the respondent's current job and uses the NS-SEC classification system. In the raw data, three variables were available representing a 5, 7, and 13-group classification. The study used the 5-group classification variable.

Missing Variables and Other Data Wrangling Issues

Cases with missing variables were omitted final sample. The study also only included cases where the child's natural mother was the primary respondent. Cases where this was not the cases were omitted. Descriptive counts for cases are reported in table 3.1. The original study did not produce such a table to compare counts with.

Table 3.1: Sample Counts for Study 1

Variables	Replicated N (Total = 4,713)
Maternal Age	
Youngest (12-19y)	209
Young (20-29y)	1,986
Middle (30-39y)	2,401
Old ($\geq$ 40y)	117
Maternal Education	
Higher degree	177
First degree	803
Diplomas in higher education	550
A/AS/S levels	546
O levels/GCSE grades A - C	1,744
GCSE grades D - G	434
Other academic qualifications	59
None of these qualifications	400
Maternal Depression	
Yes	427
No	4,286
Child Sex	
Male	2,368
Female	2,345
Maternal Parity	
No other baby	913
Yes, other babies	3,800
Maternal Socioeconomic Status	
Managerial and professional	1,936
Intermediate	790
Small employers/self-employed	631
Lower supervisory and technical	317
Semi-routine and routine	1,039
Maternal Self-Esteem	
High	4,367
Low	346
Infant Temperament	
High	2,622
Low	2,091
Maternal Psychological Distress	
High	641
Low	4,072

Original N = 4,732 / Replicated N = 4,713

3.1.2 Replicated Results and Comparisons

This original paper used multivariate OLS regression as the method of analysis with the variables listed above. It ran two models, the unadjusted model regresses SDQ TDS on Active/Withdrawal Discipline Methods without any controls. The adjusted model includes the all control variables listed in 3.1.1. This paper replicates the method and achieves the following results.

Table 3.2: Coefficients of Active Discipline Methods Variable

Active	β	p-value	95% CI
Original Unadjusted Model	1.11	< 0.01	0.93 – 1.28
Replicated Unadjusted Model	0.31	< 0.001	0.27 – 0.35
Original Adjusted Model	0.84	< 0.01	0.67 – 1.00
Replicated Adjusted Model	0.15	0.04	0.01 – 0.29

Table 3.3: Coefficients of Withdrawal Discipline Methods Variable

Withdrawal	β	p-value	95% CI
Original Unadjusted Model	0.72	< 0.01	0.53 – 0.90
Replicated Unadjusted Model	0.11	< 0.001	0.08 – 0.14
Original Adjusted Model	0.41	< 0.01	0.24 – 0.58
Replicated Adjusted Model	0.04	0.59	-0.10 – 0.17

For all four cases, the key β coefficients of interest were much smaller than what was reported in the original study. The unadjusted model's coefficients were also more statistically significant than what was produced originally.

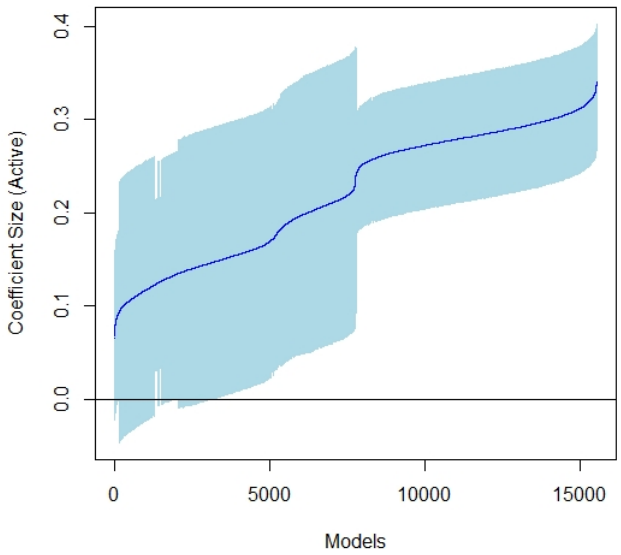
As tables 3.2 and 3.3 show, the replicated results are not close to the original study's results. A likely reason for this may include the exclusion of the child-reported mood measure which the original study's author said (but did not specify how) was included in the outcome variable.

3.1.3 Extensions and Specification Curves

A nested for loop was created and all different combinations of control variables (including their omission from the regression model) were iterated over the same base regression model. For example, a model may be run with maternal education at wave 1, at wave 4, and without maternal education at all. Another series of models may be run with all combinations of maternal education together with the 5, 7, and 13-group NS-SEC social status classifications each and also without it. This was done for all variables where a researcher degree of freedom was exploitable, resulting in about 15,552 models in total represented in figures 3.1 and 3.2 for the Active and Withdrawal coefficients respectively.

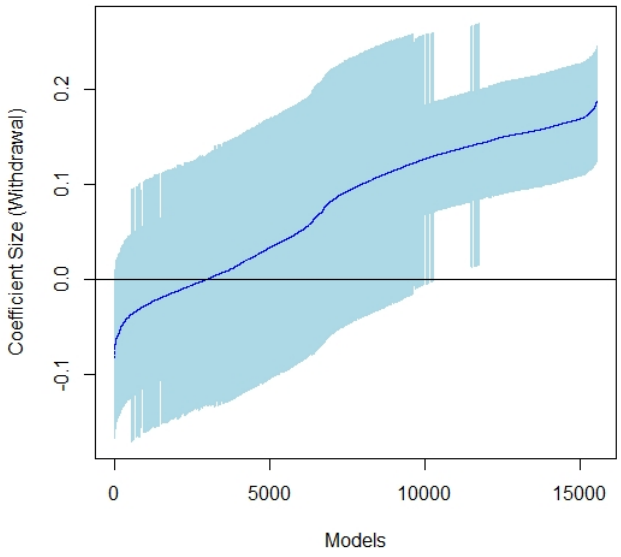
The blue line in the specification curves represents all possible values that the coefficients can take when the researcher degrees of freedom are reasonably exploited. The light blue area represents the 95% confidence interval.

Figure 3.1: Specification Curve of Active Discipline Methods Coefficient



The range of values that the Active discipline methods coefficient took range from 0.06 to 0.34. The replicated coefficient sizes of 0.31 and 0.21 for the unadjusted and adjusted versions, respectively, lie close to the upper end of the specification curve. The shape of the specification curve shows that as it is possible to 'control away' the association between the predictor and the outcome variable when many controls are added. Larger standard errors are also produced when more variables are controlled for. While there is some overlap with the 95% confidence intervals of some models with $\beta = 0$, all the coefficients of the models are above zero. However, none of the models were able to reproduce the original coefficient sizes of 1.11 or 0.84.

Figure 3.2: Specification Curve of Withdrawal Discipline Methods Coefficient



The values that the Withdrawal discipline methods coefficient took ranged from -0.08 to 0.19. The replicated coefficient value of 0.09 and 0.11 for the adjusted and unadjusted models, respectively, lie in the middle of this specification curve. There are also a large number of models where it was possible to show a negative association between Withdrawal discipline methods and

the SDQ TDS. However, there was also no combination of variables that produced the original coefficients of 0.72 and 0.41 for Withdrawal discipline methods. This paper speculates this discrepancy has to do with the exclusion of the secondary child-reported mood measure for the outcome variable.

Conditional on the range of control variables chosen above, this replication therefore finds that Active discipline methods have a decidedly positive association with a higher child socio-emotional behaviour score. Withdrawal discipline methods, on the other hand, have a more overall ambiguous relationship – it may be positive or negative depending on the variables conditioned on. This paper judges this study to have been unsuccessfully replicated.

3.2 Study 2: Zilanawala et al., 2015

Although the evidence on ethnic variations on children's socio-emotional difficulties has been well-documented (Platt, 2012), it has been decidedly mixed on the advantages and disadvantages conferred to different ethnic minority children (Zilanawala et al., 2015). Published in Social Science & Medicine, the study by Zilanawala et al. examined whether there was an ethnic gradient for child socio-emotional behaviour.

The main author of the original study, Zilanawala, was the only one of the four who responded openly and positively to attempts by this author to contact her, and an abbreviated Stata do file was received although it was not used. Correspondence with Zilanawala revealed technical discrepencies in one of the files the author had downloaded, as it was missing a large number of cases for the Index of Multiple Area Deprivation (IMD) variable. The replication has therefore omitted this (see table 3.4), but was still able to arrive at fairly close results. Unless otherwise stated, data in this study were all derived from the fourth wave of the MCS.

3.2.1 Variables and Covariates Used

Not counting the IMD variable, 13 control variables were used in the replicated multivariate OLS model with SDQ TDS as the outcome variable and ethnicity as the predictor variable. Table 3.4 contains the replicated sample size breakdowns. The original study had no such sample counts to refer to. The original N = 12,362 while the replicated N = 13,376. Categorical variables were coded and indexed beginning with 0.

Key Predictor Variables

There were three variables available in the MCS (wave 4) dataset for ***child ethnicity***, this study having chosen the 8-category typology. There was also a 6-category and an 11-category typology to choose from. This study took raw 8-category data and placed cases with 'Mixed' ethnicity into the same

group as the 'Other' category, resulting in a total of 7 categories. They are: White, Indian, Pakistani, Bangladeshi, Black Caribbean, Black African, and Other.

Outcome Variable

The ***SDQ TDS*** was used in its continuous form in this study. No re-coding was applied.

Control Variables

Child Sex or gender was re-coded to 0 for female and 1 for male.

Two variables for ***Child Age*** were available in the MCS wave 4 dataset: age in days and in years. It was not immediately apparent which the original study used, but given that it said the mean age was 7.23 years, this paper assumed age in days was used in order for such a granular figure to be arrived at. It used age in days in the replicated model.

While the original study used a 'migrant generation' variable, no such variable was directly available. Instead, in private correspondence Zilanawala said ***Mother's Year of Immigration to the UK*** was used to construct it – although it was not obvious how she arrived at a 'generation' variable from year of migration. The replication used the year of migration variable in a banded form.

Maternal Highest Educational Qualifications was asked at wave 4 and used the NVQ equivalence scale with seven possible responses: NVQ levels 1 to 5, Overseas Qualifications Only, and None of These. They were re-coded to take values 0 to 6 respectively.

Maternal Employment was used in the original study as three having three categories: full-time work, part-time work, and not in work. However, the variable containing information on part-time work had too many missing values and could not be appropriately used. A separate variable asking respondents whether they were in work or not was used in the replication as

a binary variable instead.

Log of Income was created by taking logarithm of the equivalised continuous household income variable in the wave 4 dataset.

Whether English was the main language used at home was used as a binary variable, re-coded to 1 and 0 for Yes/No responses respectively.

Frequency of racism experienced was coded from a question asking respondents about the frequency that they had experienced racist attacks or insults. Four options were available: Not at all, Not very common, Fairly common, and Very common. The first two responses were coded 0 and the last two coded 1 to create binary variable.

The question asking ***Whether child has regular bedtime*** at wave 4 had four responses: Always, Usually, Sometimes, or Never. The first two were re-coded into 1 and the last two into 0 to make a binary variable.

Home Learning Environment referred to the sum of coded scores of four questions: how often a mother reads to her child, and whether she helps her child with reading, writing, and maths. The last three questions were re-coded 1 and 0 for yes/no responses. The first question had 7 possible responses originally coded 1 to 6 for frequencies 'Every day or almost every day', 'Several times a week', 'Once or twice a week', 'Once or twice a month', 'Less often than once a month', and 'Not at all'. The last two of these responses were combined and re-coded into 0 and the rest into 1. The resulting four binary variables were then summed as a way to count the number of activities a mother helps her child in.

Discipline Strategies Score was derived from questions in wave 4 originating from the Conflict Tactics Scale (Straus, 2017). Questions asked to mothers include 'how often do you ignore/smack/shout/send to bedroom or naughty chair/take away treats/tell off/bribe with sweets or other ... when [child] is naughty.' Frequencies were coded on a 6-point scale – categories: can't say, never, rarely, once a month, at least once a week, and daily –

labelled 0 to 5 respectively before being summed into a continuous variable.

Maternal Distress was assessed at wave 4 with the Kessler 6 Scale (K6) (Kessler et al., 2002). The resulting score was available in the raw dataset as a continuous variable and was used without re-coding.

Parental Basic Skills Difficulties was comprised of responses from three questions on whether they could: read a children's book, fill in forms, and check change in a shop. This was asked at MCS wave 1. There were three possible responses for each question: 'Yes, easily', 'Yes, with difficulty', and 'No', coded 0 to 2 respectively. These were then summed with a higher score (from 0 to 6) indicating greater difficulty. This was used as a categorical variable in the replication.

Missing Variables and Other Data Wrangling Issues

Cases of children diagnosed with ***ADHD*** or ***Asperger's*** at wave 4 were omitted from the sample.

The variable ***Index of Multiple Area Deprivation (IMD)*** was used in the original study, constructed out of individual questions given to respondents relating to area deprivation separately in England, Scotland, Wales, and Northern Ireland. The downloaded dataset only had England and Northern Ireland data. The author was unable to resolve this issue and so omitted this variable.

For the control variables with missing values, the original study filled missing values with multiple imputation using Stata command 'mi'. The replication also used multiple imputation but used the R package 'Hmisc'[1]. Since the same software was not used compared to the original study, which analysed the data in Stata, this paper acknowledges that the imputed values may not be the same.

[1]See https://CRAN.R-project.org/package=Hmisc for details

Table 3.4: Sample Counts for Study 2

Variables	Replicated N
Ethnicities	
White (Reference Group)	11,382
Indian	314
Pakistani	527
Bangladeshi	198
Black Caribbean	151
Black African	229
Other	561
Child Sex	
Male	6,793
Female	6,569
Year of Immigration to UK (Mother)	
Non-migrant	11,813
1960 to 1969	104
1970 to 1979	237
1980 to 1989	332
1990 to 1999	776
2000 and later	100
Maternal Educational Qualification	
NVQ level 1	930
NVQ level 2	3,580
NVQ level 3	2,063
NVQ level 4	4,145
NVQ level 5	932
Overseas qualifications	344
None of these qualifications	1,368
Is English Main Language at Home	
Yes	12,415
No	947
Maternal Employment	
Not in work	4,906
In work	8,456
Racism in Area is	
Very common	159
Fairly common	2,845
Not common	10,358

Original N = 12,376 / Replicated N = 13,362 (with multiple imputation)

Sample Counts for Study 2 (Continued)

Variables	Replicated N
Child Regular Bedtime	
Sometimes/Always	12,119
Never/Rarely	1,243
Home Learning Environment	
Receives no help	481
Receives help in 1 area	2,667
Receives help in 2 areas	2,014
Receives help in 3 areas	2,806
Receives help in all areas	5,394
Parental Basic Skills Difficulties	
Difficulty score 0	12,100
Difficulty score 1	295
Difficulty score 2	515
Difficulty score 3	70
Difficulty score 4	132
Difficulty score 5	18
Difficulty score 6	232
Index of Multiple Area Deprivation[a]	
Quintile 1	N.A.
Quintile 2	N.A.
Quintile 3	N.A.
Quintile 4	N.A.
Quintile 5	N.A.

Variables	Replicated Mean (Std Dev.)
Child Age (Days)	2,640 (89.94)
Maternal Distress Score	4.15 (3.87)
Log of Income	5.78 (0.63)
Discipline Strategies Score	17.84 (4.00)

Original N = 12,376 / Replicated N = 13,362 (with multiple imputation)

[a] Data not available to replication study.

3.2.2 Replicated Results and Comparisons

The original study used multivariate OLS regression as the method of analysis. It ran two models. The base model regressed the SDQ TDS on ethnicities with only age and sex as controls. The controlled model had the same out-

come and predictor variable but included all controls in the above section. The resulting coefficients are listed in table 3.5.

Table 3.5: Coefficients of Ethnicities

Ethnicity	Original β (SE)	Replicated β (SE)
Base Model		
Indian	-0.037 (0.39)	0.58 (0.31)
Pakistani	2.43*** (0.33)	2.11*** (0.24)
Bangladeshi	1.60*** (0.40)	1.81*** (0.39)
Black Caribbean	1.71*** (0.31)	1.04* (0.44)
Black African	-0.23 (0.42)	-1.08** (0.36)
Other	-0.11 (0.42)	0.90*** (0.24)
Controlled Model		
Indian	0.21 (0.37)	0.34 (0.28)
Pakistani	0.69* (0.31)	0.30 (0.24)
Bangladeshi	-0.023 (0.46)	-0.37 (0.37)
Black Caribbean	0.61 (0.36)	0.44 (0.38)
Black African	-1.29*** (0.35)	-1.82*** (0.33)
Other	-0.70 (0.37)	0.078 (0.21)

*** < 0.001 ** < 0.01 * < 0.05 . < 0.1

This replication showed mixed results, with not all of the coefficients close to the original. However, the coefficients for a number of categories were close – such as the Pakistani and Bangladeshi coefficient for the base model, and the Black African coefficient in the controlled model (in magnitude, direction, and significance). However, notable differences between the original study and the replication included the base model's Black Caribbean coefficient and the controlled model's Pakistani coefficient. The most striking difference was in the Other coefficient given how neither the direction, magnitude, nor significance for either model was close. Yet, this replication was on a whole successful because it was mostly able to reproduce the coefficients which were originally significant. This paper speculates that differences were attributable to the missing area deprivation variable and the differing imputation software used.

3.2.3 Extensions and Specification Curves

Specification curves were created by building a nested for loop iterating over every combination of control variables (including combinations with their omission) and then extracting the coefficients for all of the ethncities. This resulted in 16,384 models. The blue lines represent all possible values the coefficients took and the light blue regions represents the 95% confidence interval.

This is shown in figures 3.3 to 3.8, with the red dot representing the original study's base model coefficient and the brown dot representing the original study's controlled model coefficient. Note that individual coefficients were sorted in ascending order – i.e. model 10 for Indian coefficient is not necessarily the same as model 10 for Pakistani coefficient etc.

The specification curve for the Indian coefficient ranged from -0.02 to 1.33. The study's coefficients of -0.037 (base) and 0.21 (controlled) appeared on the bottom tail of the specification curve, indicated by the red and brown dot respectively in figure 3.3. The Black African ethnicity's original coefficients of -0.23 and -1.29 also featured on a tail end of the specification curve (-0.23 is not on the blue line but falls within the 95% confidence interval). Most of the other ethnicities had original coefficients either much closer to the middle of the curve, or whose original base and controlled model values spanned a large distance over the specification curve. All of the coefficients in the original study featured within the 95% confidence intervals of their respective replicated specification curves.

Figure 3.3: Specification Curve of Indian Ethnicity Coefficient

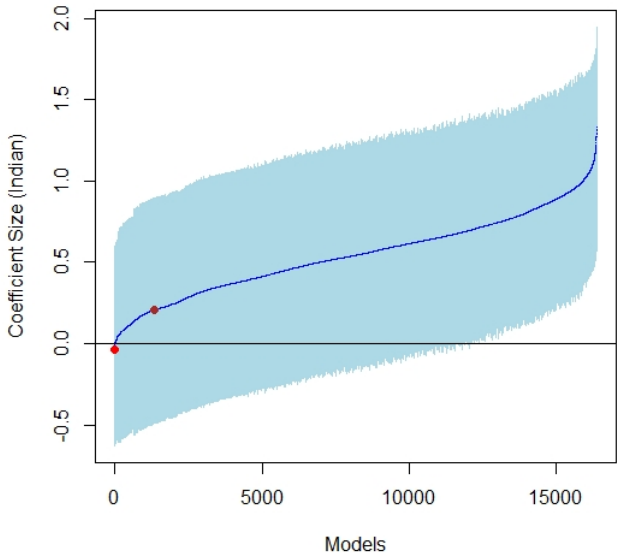


Figure 3.4: Specification Curve of Pakistani Ethnicity Coefficient

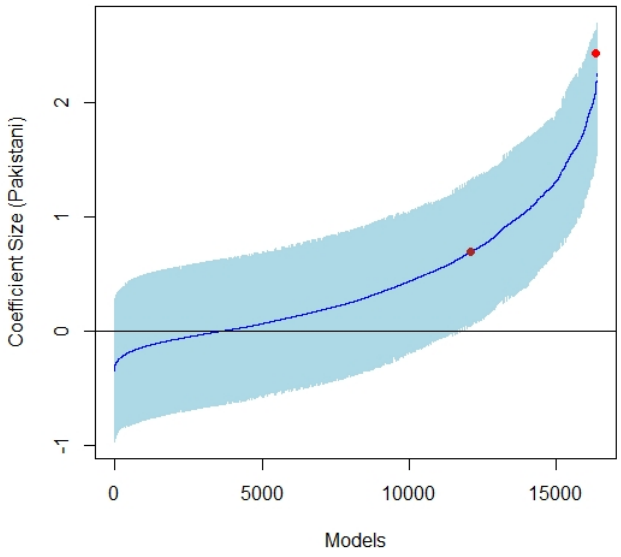


Figure 3.5: Specification Curve of Bangladeshi Ethnicity Coefficient

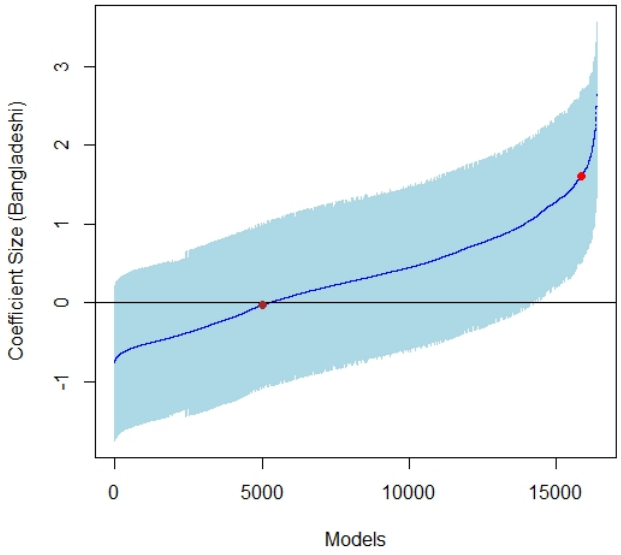


Figure 3.6: Specification Curve of Black Caribbean Ethnicity Coefficient

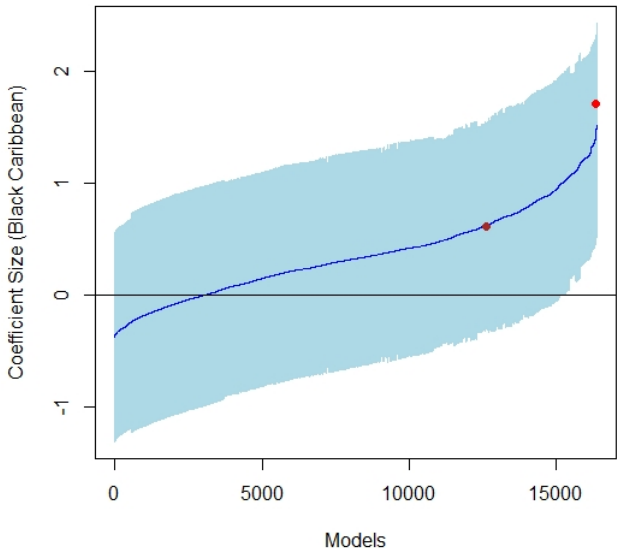


Figure 3.7: Specification Curve of Black African Ethnicity Coefficient

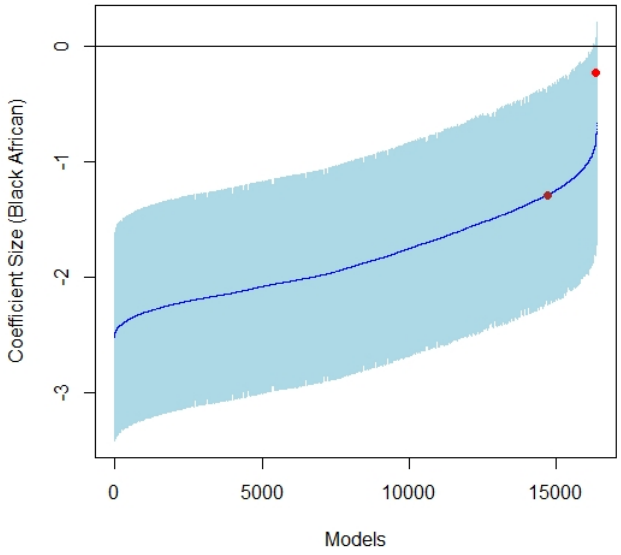
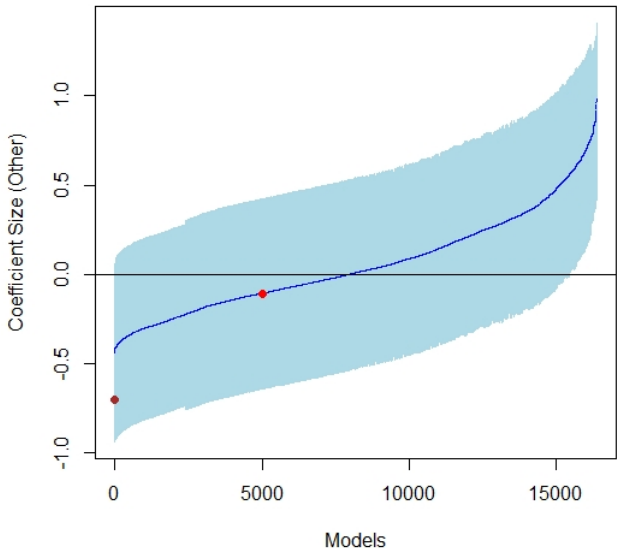


Figure 3.8: Specification Curve of Other Ethnicity Coefficient



Some of the original study's coefficients could not be reproduced within their specific replicated specification curves. The Pakistani specification curve had coefficients take values from -0.36 to 2.25, but the original study's base model had a coefficient value of 2.43. The Black Caribbean specification curve had its coefficients take values from -0.38 to 1.51, but the original study's base model had a coefficient value of 1.71. The most dissimilar was the Black African specification curve, where it took values of -2.53 to -0.68 but the original study's base model had a coefficient of -0.23. However, all three of these examples still had their original coefficients within the 95% confidence interval of at least one replicated model. On balance, this study has been successfully replicated.

With the exception of the Black African coefficient, all of the specification curves cross the $\beta = 0$ line which indicate that it is possible to both control away the association and show opposing directional associations between a given ethnicity and child socio-emotional behaviour.

This is most apparent for the Other coefficient. Of the 16,384 models ran, 8,043 (just under half) had a negative coefficient value which indicates that there are about just as many ways to show a positive and negative association between this category and child socio-emotional development. This, however, makes theoretical sense given how 'Other' is not a coherent category and so we may expect its association to fluctuate with respect to conditioning on different control variables. What was notable was how the original study showed this coefficient to only be in the negative region. There was also no possible combination of variables to reproduce the original study's controlled model coefficient.

The final point to make is that, except for the Pakistani specification curve, all of the models have relatively flat specification curves. This indicates that, with respect to the list of control variables applied, the coefficients are generally not too sensitive to researcher degrees of freedom. The Pakistani coefficient was the exception as it tails relatively sharply upwards in the

rightmost region of the curve.

Like the original study, this replication finds evidence for an ethnic gradient for child socio-emotional behaviour scores. Conditional on the above list of control variables, the Indian ethnicity had almost all replicated specifications result in a positive association between the SDQ TDS. The Black African ethnicity, on the other hand, had all of the replicated models show a decidedly negative association. Thus, compared to white children, Indian children had higher average behavioural problem scores while Black African children seemed to have lower average behavioural scores. Most other ethnicities – Pakistani, Bangladeshi, Black Caribbean – also had an overall positive association with the outcome variable, though it was possible for these three groups to also result in a negative association. The Other ethnicity category had the most ambiguous relationship, with about half of the models producing a positive and half a negative association.

3.3 Study 3: Noonan et al., 2018

Building on well-documented associations between household or family income and socio-emotional behaviour in early childhood, using MCS wave 5 data (child age 11) this study examines whether this relationship holds in later childhood as well (Noonan et al., 2018). Specifically, its key predictor variable was permanent family income, which took an average of the logarithm of income over the first five waves. This was published in SSM – Population Health, which shares the same editorial board[2] as its sister journal Social Science & Medicine.

3.3.1 Variables and Covariates Used

This was the most extensively controlled study replicated, with 25 control variables used. Permanent family income was the key predictor variable with a binary SDQ TDS used as the outcome variable. Table 3.6 contains the original and replicated sample counts. The original study's final N = 8,499 and the replicated N = 9,034 after omitting cases with missing values. All data came from the MCS wave 5 survey unless otherwise stated. Categorical variables were again coded and indexed beginning with 0.

Key Predictor Variables

Permanent Family Income was stated in the study to have been calculated by 'averaging equivalised household income over the first five surveys, expressed in logarithmic form' (Noonan et al., 2018). It was not immediately clear whether the study had (1) firstly taken an arithmetic mean of continuous household income over the first five waves and secondly taking its logarithm, or (2) firstly taken logarithm of continuous household at each wave before secondly finding the arithmetic mean. Methods (1) and (2) result in different figures. The replication used method (2) in the initial analysis but used both methods in the multiverse analysis.

[2]See https://www.journals.elsevier.com/ssm-population-health

Outcome Variable

SDQ TDS was used as a binary variable, with a raw score of 17 or more indicating problematic socio-emotional behaviour (coded 1) and under 17 indicating normal behavioural levels (coded 0).

Control Variables

Socio-demographic Variables

Child Age was given in years, with three possible ages: 10, 11, or 12.

Child Sex was binary: male or female (coded 1 and 0 respectively).

Child Ethnicity was also used as a binary variable, white or non-white (coded 0 and 1). In this case, it did not matter which of the 3 original classification variables were chosen as the numbers of white and non-white children were the same. The multiverse analysis included the three ethnicity variables with 6, 8, and 11-category variables.

The ***Number of Siblings*** of the child was asked at MCS wave 4 with three categories: no siblings, one sibling, two siblings, or three and more siblings, coded 0, 1, 2, and 3 respectively.

Data for highest ***Maternal Educational Qualifications*** was gathered from MCS wave 1 and coded into four categories: Tertiary degree, A-levels, GCSE, or None of these (coded 0 to 3 respectively).

Maternal Age in Years at Childbirth in the original study was re-coded from a continuous variable into the following groups: $<$20, 20-24, 25-29, 30-34, and $\geq$35. An existing banded age group was available (categories: 12-19, 20-29, 30-39, $\geq$40) and was included in the multiverse analysis, along with age as a continuous variable.

Housing Tenancy Status was asked at MCS wave 4 and then re-coded into three categories: own/mortgage, council housing, and rent/other.

Child Health Endowment Variables

Child Birthweight (at MCS wave 1) was a binary, created by splitting the original continuous variable of birthweight in kg into two groups: >2.5 kg and ≤2.5 kg.

Child Gestational Age (at MCS wave 1) was given in days in its raw form. It was converted to weeks and split into two groups to be used as a binary variable: <37 weeks and ≥37 weeks.

Whether child has longstanding illness was asked at MCS wave 4 with yes and no responses coded into 2 groups.

Mother's drinking behaviour during pregnancy (at MCS wave 1) was re-coded from a 7-category frequency count to a 4-category one: Never, Light, Moderate, and Heavy/binge.

Mother's smoking behaviour during pregnancy (at MCS wave 1) had frequency counts in its raw form, but was re-coded into three groups: Never smoked, Stopped smoking during pregnancy, and Smoked throughout pregnancy. The original frequency counts were collapsed into the third option.

Mother's breastfeeding duration (at MCS wave 1) was coded into five categories: no breastfeeding, < 7 days, 1 week to 3 months inclusive, 3 to 6 months inclusive, and ≥ 6 months. The original data was available in a more granular form which were collapsed into these groups.

Mother's psychological distress pattern used individual psychological screening tests – the Rutter Malaise Inventory (RMI) (Rutter, Tizard, & Whitmore, 1970) was used in MCS wave 1, while the Kessler 6 Scale (K6) (Kessler et al., 2002) was used in all other waves – to measure maternal distress. Each screening's results were converted into a binary score by summing their individual components and dividing it along the pre-defined cutoffs defined in their respective literature. For the RMI, a score of ≥ 4 was coded 1 while < 4 was 0. For the K6, ≥ 6 was coded 1 while < 6 was 0.

All the individual binary markers at each wave were then used to create 7 categories for the final variable used: Never, Early (MCS1 and/or 2), Middle (MCS3 and/or 4), Current (MCS wave 5), Recurrent (MCS 1 and any other wave), Persistent (all waves), and Others. This information was taken from the original study.

Family/External Factors Variables

All variables listed below are derived from MCS wave 4 data.

Change in mother's relationship status from waves 4 to 5 was constructed from a question asking mothers about her relationship status in waves 4 and 5, and then comparing the responses (single or partnered) that she gave across waves. If both responses were the same, this was classified as having no change in the relationship. If she became partnered or became single, these were also recorded.

Whether child has regular bedtimes on weekdays was directly available and used in its raw form with four categories: Never/almost Never, Sometimes, Usually, and Always coded in four categories.

Hours child spends playing computer/video games during a week was also used in its raw form, with four categories: None, Less than an hour, 1 to less than 3 hours, and $\geq$ 3 hours.

Mother's satisfaction with time spent with child was available in a more granular form in the MCS wave 4 data. It was re-coded into three categories: Enough/more than enough, Just enough, and Not enough.

Whether a ***Mother tells child off when naughty*** was one of three discipline-related questions used in the original study. This was coded into four categories: Never, Rarely, Sometimes, and Often.

Whether the ***Mother sends child to room when naughty*** and if ***she takes things away from child when naughty*** was used and coded in the same manner as the above question on telling her child off. The replication

added the scores of all three questions together to generate a composite score on child discipline. This was used in the extended multiverse analysis.

How often mother reads to child was originally coded into 6 categories of differing frequencies, re-coded into categories: Not at all, Approximately monthly, and Weekly or more. ***How often mother plays with her child*** was available and treated in the same way.

Whether mother has longstanding health issues was answered with yes and no responses and coded into a binary variable.

Whether mother currently smokes was derived from an original question asking for the frequency of her smoking, but was then re-coded into Non-smoker and Smoker categories.

Mother's current alcohol intake was available as a seven-category measure but was re-coded into frequencies: Never/almost never, Monthly, 1 to 2 times weekly, 3 to 4 times weekly, and $\geq$5 times weekly.

How often does child spend time with friends outside school was coded into the following categories: Not at all, Approximately monthly, Regularly 1 to 3 times a week, and Most days.

Table 3.6: Sample Counts for Study 3

Variables	Original N	Replicated N
Socio-demographic Variables		
Age		
10	2,800	2,951
11	5,666	6,049
12	33	34
Child Sex		
Male	4,260	4,511
Female	4,239	4,523
Ethnicity		
White	7,650	8,053
Non-white	849	981
Siblings 0	968	1,017
1	4,041	4,260
2	2,381	2,557
≥ 3	1,109	1,200
Mother's Academic Qualification		
Tertiary degree	2,599	2,733
A-levels	926	978
GCSE	3,994	4,058
None of these	980	1,265
Mother's Age at Childbirth		
< 20 years	507	538
20 – 24 years	1,260	1,383
25 – 29 years	2,410	2,561
30 – 34 years	2,797	2,961
≥ 35 years	1,516	1,591
Housing Tenure		
Own/Mortgage	6,177	6,560
Council	938	818
Rent/Other	1,384	1,656
Child Health Endowment Variables		
Birthweight		
> 2.5 kg	8,009	8,519
≤ 2.5 kg	490	515
Gestational Age		
≥ 37 weeks	7,886	8,449
< 37 weeks	613	585

Original N = 8,499 / Replicated N = 9,034

Sample Counts for Study 3 (Continued)

Variables	Original N	Replicated N
Child Longstanding Illness		
Yes	1,571	1,659
No	6,928	7,375
Mother's Alcohol Intake In Pregnancy		
Never	5,715	6,085
Light	2,164	2,076
Moderate	449	697
Heavy/binge	171	176
Mother's Smoking In Pregnancy		
Never smoked	5,690	6,113
Stopped smoking during pregnancy	1,076	1,116
Smoked throughout pregnancy	1,733	1,805
Breastfeeding Duration		
No breastfeeding	2,482	2,566
$<$ 7 days	979	631
1 week to 3 months (inclusive)	2,127	1,209
3 to 6 months (inclusive)	1,225	1,480
$>$ 6 months	1,686	3,148
Maternal Psychological Distress Pattern		
Never	3,185	2,837
Early years only	520	657
Middle years only	604	647
Age 11 only	637	940
Other pattern	1,237	1,398
Recurrent	1,930	2,247
Persistent	386	308
Family/External Factors Variables		
Change in Mother's Relationship, wave 4 – 5		
Became single	531	652
Became partnered	358	377
No change	7,610	8,005
Regular Bedtime on Weekdays		
Never/almost never	288	312
Sometimes	435	467
Usually	2,699	2,854
Always	5,077	5,401

Original N = 8,499 / Replicated N = 9,034

Sample Counts for Study 3 (Continued)

Variables	Original N	Replicated N
Hours spent playing computer/video games during week		
None	866	936
$<$ 1 hour	4,644	4,901
1 – 3 hours	2,668	2,848
$>$ 3 hours	321	349
Mother's satisfaction with time spent with child		
Enough or more than enough	2,029	2,177
Just enough	3,682	3,902
Not enough	2,788	2,955
Tells off when naughty		
Never	48	59
Rarely	1,019	1,117
Sometimes	3,003	3,199
Often	4,429	4,659
Sends to room when naughty		
Never	1,037	1,109
Rarely	2,498	2,668
Sometimes	3,406	3,635
Often	1,558	1,622
Takes things away when naughty		
Never	687	750
Rarely	2,414	2,576
Sometimes	3,931	4,162
Often	1,467	1,546
Amount of time reading to child		
Not at all	176	195
Approximately monthly	615	657
Weekly or more	7,708	8,182
Time spent playing games with child		
Not at all	385	417
Approximately monthly	2,226	2,366
Weekly or more	5,888	6,251
Mother longstanding health condition		
Yes	2,117	2,263
No	6,382	6,771
Mother currently smoking		
Smoker	2,116	2,280
Non-smoker	6,383	6,754

Original N = 8,499 / Replicated N = 9,034

Sample Counts for Study 3 (Continued)

Variables	Original N	Replicated N
Mother's current alcohol intake		
$\geq$ 5 times weekly	576	611
3 – 4 times weekly	1,069	1,131
1 – 2 times weekly	2,445	2,587
Monthly	3,067	3,096
Never/almost never	1,342	1,609
Time child spent with friends outside school		
Not at all	440	491
Approximately monthly	1,532	1,652
Regularly, 1 – 3 times per week	4,667	4,911
Most days	1,860	1,980
Time child spends playing sport		
Never/rarely	2,334	2,546
Once per week	2,309	2,420
2 – 3 times per week	3,154	3,323
$\geq$ 4 times per week	702	745

Original N = 8,499 / Replicated N = 9,034

How often does child play sports was asked at wave 4 and was coded into the categories: Never/rarely, Once a week, 2 to 3 times a week, and 4 or more times a week.

Missing Variables and Other Data Wrangling Issues

The study subsetted cases for ***Non-Singleborns*** i.e. cases of families with twin, triplet, or higher births were removed from the sample. Cases with any missing variables were also removed from the sample, resulting in a final replicated sample size of 9,034, compared to the original sample size of 8,499. Variable counts are available in table 3.6 above. Most of the variables had categories with counts of a similar frequency.

3.3.2 Replicated Result and Comparison

This paper used multivariate logistic regression and reported marginal effects derived from Stata. This study used the method and R code found in a working paper (Fernihough, 2011) (available online[3]) to replicate this. There were direct comparisons of their method with the corresponding Stata command for calculating marginal effects (command: mfx) which returned similar outputs. One limitation is that the paper does not provide code to produce average marginal effects which is what Stata calculates by default. The paper instead gives code which produces the average of the sample marginal effects and their associated standard errors. This is similar enough for the purposes of this paper but it is not the equivalent metric produced.

Re-running the same model yielded results presented in table 3.7. The original study had no results for a base model without any controls.

Table 3.7: Marginal Effect of Family Income (Fully Controlled)

Family Income	Original ME (SE)	Replicated ME (SE)
log of income	-0.044*** (0.014)	-0.040** (0.011)

*** p < 0.01 ** p < 0.05 * p < 0.1

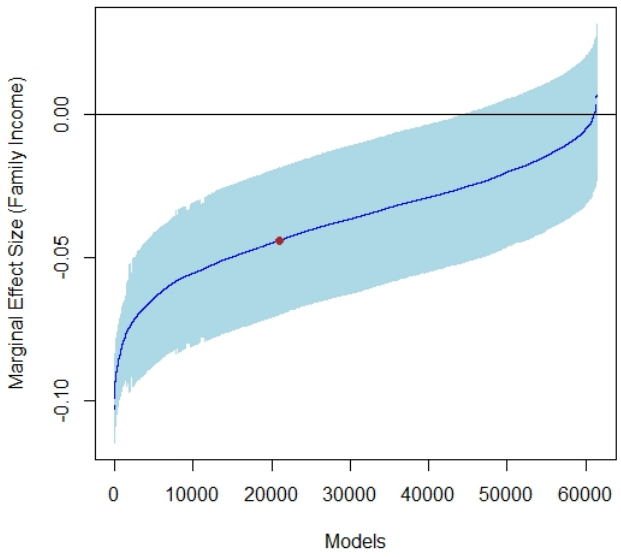
While not as statistically significant as the original study, the direction and magnitude of the derived marginal effect was the closest of all four studies replicated. This shows a negative relationship between log income and child socio-emotional behaviour. The higher the permanent family income, the less likely it is for a child to have a problematic behavioural level score.

[3]See http://www.ucd.ie/t4cms/WP11_22.pdf

3.3.3 Extension and Specification Curve

As in previous studies, a nested for loop was built which iterated over every combination of control variables and its variants listed in section 3.2.1 before extracting the relevant coefficients and calculating the marginal effects from it. This resulted in 61,440 models, the most of any of the replications done in this paper. This was due to the large number of control variables this study had used. The specification curve is presented in figure 3.9.

Figure 3.9: Specification Curve of Family Income Marginal Effect



The blue line represents all possible marginal effect values after exhausting researcher degrees of freedom, and the light blue region represents the 95% confidence interval. The figure above shows that the coefficients can take values from -0.103 to 0.007 conditioning on different sets of control variables. Most of the coefficients show a negative association, with only 311 of the 61,440 models showing a positive relationship.

Apart from the at the lower end of the specification curve (the bottom end

of the curve likely represents results from base/uncontrolled models), the marginal effect of log permanent family income does not seem to be too sensitive to the researcher degrees of freedom in variable selection and inclusion, indicating a relatively robust negative association.

The original study's marginal effect of -0.044 lies close to the middle of the specification curve, which hints that there was no specification searching on the part of the original authors. On the other hand, there were a number of models whose 95% confidence intervals (and some models whose marginal effects) crossed the $ME = 0$ line, indicating that it was again possible to 'control away' the marginal effect of family income.

This paper judges this study to have been replicated very successfully and whose marginal effects are robust.

3.4 Study 4: McMunn et al., 2012

Published in the Journal of Epidemiology and Community Health, which falls under the BMJ Group, this study examines the association between maternal employment and child socio-emotional behaviour. It cites the mixed evidence presented on child outcomes as mothers increasingly put more of their time in the workplace and away from the home (McMunn et al., 2012). While it stated that it sought to examine the 'effect' of maternal employment and parental working arrangements on child socio-emotional behaviour, its causal language may be misleading as its multivariate logit model does not imply a causal connection but only an associative one. It is this which this paper replicates.

3.4.1 Variables and Covariates Used

The study had 6 control variables, the fewest of the four studies replicated. For the multiverse analysis, the replication included some variables used in other studies. They include permanent income, child ethnicity, child age, and the 'child health endowment variables' used in Study 3: birthweight, gestational age, and breastfeeding data, as well as the mother's alcohol consumption and smoking behaviour during pregnancy. Categorical variables were again coded and indexed beginning with 0.

Key Predictor Variables

The ***Maternal Employment*** variable was constructed from MCS waves 1, 2, and 3 data. At each wave mothers were asked if they were (coded 1) or were not (coded 0) in employment. Mothers who were at home on maternity leave at wave 1 was considered to be employed, as in the original study. The 1/0 values were then summed as a way of counting the number of waves that mothers were employed. Mothers employed at no waves scored 0, one wave scored 1, two waves scored 2, and three waves scored 3. This four-category variable was used in the analysis.

Outcome Variable

SDQ TDS was used in its binary form, with a score of 17 and above representing abnormally high levels of socio-emotional behaviour, and a score below 17 indicating normal behavioural levels.

Control Variables

Paternal Employment was calculated in the same way as for maternal employment, above, with a four-category variable reflecting the number of waves that fathers were employed. However, a fifth category was added to reflect cases in which there was an absent father in at least one wave. Cases with absent fathers were not omitted.

The highest level of ***Maternal Education*** was available in MCS wave 3 and used the NVQ equivalence scale with seven possible responses: NVQ levels 1 to 5, Overseas qualifications only, and None of these.

Mother's Age at Childbirth in years was available at wave 1 and was used in its raw form. A separate banded age variable was also available and was included in the multiverse analysis.

Household Income (Banded) was constructed by combining household income bands over the first three waves. Even though different income bands were used at each wave, the original study said it broke each wave's data down to its quintiles and combined each case's quintile position over the three waves to arrive at a final score. The replication used the same approach but could not reproduce the original study's quintiles. Instead, it could only reasonably break down income data from each wave into a 6-band distribution (as the original banded data did not seem to come in a granular-enough format) and combined it instead.

Continuous household income was used in the extended multiverse analysis. Two forms of it were included: taking logarithm of the arithmetic mean of income over 3 waves, and taking the arithmetic mean of log income in each of the 3 waves.

Maternal psychological distress was a 4-category variable constructed in a similar way as maternal employment. It used individual psychological screening tests – the Rutter Malaise Inventory (RMI) (Rutter et al., 1970) was used in MCS wave 1, while the Kessler 6 Scale (K6) (Kessler et al., 2002) was used in the two other waves – to measure maternal distress. Each wave's results was converted into a binary score by summing all their individual components and dividing it along the pre-defined cutoffs as defined in their respective literatures. For the RMI, a score of $\geq$4 was coded 1 while $<$4 was 0. For the K6, $\geq$6 was coded 1 while $<$6 was 0. Individual binary markers at each wave were then used to create 4 categories for the final variable used: Distressed at no sweeps (scored 0), Distressed for one sweep (scored 1), Distressed for two sweeps (scored 2), and Distressed for three sweeps (scored 3).

For Specification Curve Construction Only:

Child Age in years at MCS wave 3 was available in both a continuous and banded form. Both were used.

All pregnancy-related data was available in wave 1. ***Child Birthweight*** was included as a binary variable: weights $>$2.5kg and $\leq$2.5kg.

Child Gestational Age was given in days and was converted to weeks. A binary variable was used: gestational age $\geq$37 weeks and $<$37 weeks.

The extent to which the ***Mother was drinking during pregnancy*** was coded into 4 categories: Never, Light, Moderate, and Heavy/binge.

Mother smoking during pregnancy was coded into 3 categories: Never smoked, Stopped smoking during pregnancy, and Smoked throughout pregnancy.

Maternal breastfeeding duration was used as a 5-category variable: No breastfeeding, less than 7 days, 1 week to 3 months inclusive, 3 to 6 months inclusive, and more than 6 months.

Missing Variables and Other Data Wrangling Issues

Cases of ***Non-Singleborns*** (births of twins, triplets, and above) were omitted from the sample, as were cases with ***Absent Mothers*** and ***Non-white Children***. However, child ethnicity was factored in separately in the multiverse analysis. All other cases with missing variables were omitted.

Child Sex or gender was not explicitly modelled but was used to separate the main sample into two sub-samples which were analysed separately.

Variable lists and sample counts/descriptive statistics are found in table 3.8 below. Figures for boys and girls are presented separately.

Table 3.8: Sample Counts for Study 4 (Boys)

Variables	Replicated N
Mother's Employment	
Paid work 3 sweeps	1,244
Paid work 2 sweeps	847
Paid work 1 sweep	1,012
Paid work no sweeps	1,688
Father's Employment	
Paid work 3 sweeps	2,376
Paid work 2 sweeps	265
Paid work 1 sweep	75
Paid work no sweeps	68
Father not present	2,007
Mother's Education	
NVQ level 1	349
NVQ level 2	1,320
NVQ level 3	785
NVQ level 4	1,635
NVQ level 5	263
Overseas qualification only	89
None of these	350
Maternal Distress	
Distressed all 3 sweeps	222
Distressed for 2 sweeps	437
Distressed for 1 sweep	871
Distressed at no sweeps	3,261

Variables	Mean (Standard Deviation)
Mother's Age	
Age at childbirth in years	28.83 (5.78)
Household Income Band	
Mean income band score	7.25 (3.65)

Original N = 5,032 / Replicated N = 4,791

Table 3.9: Sample Counts for Study 4 (Girls)

Variables	Replicated N
Mother's Employment	
Paid work 3 sweeps	1,210
Paid work 2 sweeps	733
Paid work 1 sweep	984
Paid work no sweeps	1,661
Father's Employment	
Paid work 3 sweeps	2,253
Paid work 2 sweeps	275
Paid work 1 sweep	78
Paid work no sweeps	80
Father not present	1,902
Mother's Education	
NVQ level 1	319
NVQ level 2	1,310
NVQ level 3	722
NVQ level 4	1,556
NVQ level 5	294
Overseas qualification only	65
None of these	322
Maternal Distress	
Distressed all 3 sweeps	190
Distressed for 2 sweeps	401
Distressed for 1 sweep	782
Distressed at no sweeps	3,215

Variables	Mean (Standard Deviation)
Mother's Age	
Age at childbirth (years)	28.91 (5.77)
Household Income Band	
Mean income band score	7.28 (3.64)

Original N = 4,472 / Replicated N = 4,588

3.4.2 Replicated Results and Comparisons

The original paper used multivariate logistic regression and took exponents of the coefficients to arrive at Odds Ratios (ORs). The replication used the same approach and the results are presented in tables 3.10 and 3.11. Cases with mothers employed in all three waves were used as the reference group. The original study did not report p-values and only showed the 95% confidence intervals; the replication followed suit.

Table 3.10: Odds Ratio of Maternal Employment (Boys)

In paid work	Original OR (95 % CI)	Replicated OR (95 % CI)
Unadjusted		
All three waves	1.00	1.00
Two waves	1.15 (0.85 to 1.56)	1.36 (0.97 to 1.91)
One wave	1.58 (1.20 to 2.08)	2.01 (1.46 to 2.77)
No waves	2.29 (1.81 to 2.88)	3.52 (2.71 to 4.63)
Adjusted		
All three waves	1.00	1.00
Two waves	0.93 (0.67 to 1.28)	0.98 (0.64 to 1.49)
One wave	0.94 (0.69 to 1.27)	1.11 (0.73 to 1.67)
No waves	1.00 (0.76 to 1.32)	1.20 (0.81 to 1.78)

Table 3.11: Odds Ratio of Maternal Employment (Girls)

In paid work	Original OR (95 % CI)	Replicated OR (95 % CI)
Unadjusted		
All three waves	1.00	1.00
Two waves	2.27 (1.49 to 3.45)	2.74 (1.66 to 4.62)
One wave	2.84 (1.88 to 4.31)	3.28 (1.97 to 5.56)
No waves	6.00 (4.29 to 8.37)	9.24 (6.13 to 14.56)
Adjusted		
All three waves	1.00	1.00
Two waves	1.49 (0.96 to 2.31)	1.37 (0.75 to 2.51)
One wave	1.39 (0.88 to 2.18)	1.19 (0.64 to 2.23)
No waves	2.01 (1.35 to 2.97)	1.74 (1.00 to 3.10)

The replicated results are on balance in line with the original results. Most of the replicated odds lie within the 95% confidence interval of the original odds ratios. The exceptions to this include the unadjusted model for both boys and girls comparing No waves to All three waves. For boys, the replicated odds of 3.52 lies outside the 1.81 to 2.88 band. For girls, the replicated odds of 9.24 lies outside the 4.29 to 8.37 band. For boys whose mothers were in paid work for one wave in the adjusted model, the original study showed reduced odds (<1) for socio-emotional problems, but the replication showed increased odds (>1) instead. The same applies for the respective No waves categories also, showing equal odds for behavioural problems but the replication showing greater odds.

For both boys and girls, the replication confirms the finding that the odds of a child having problematic socio-emotional behaviour increases the more waves their mother is unemployed. For the same set of control variables, the odds are also higher for girls than for boys. Maternal employment seems to be more heavily associated with girls' behavioural problem than for boys.

3.4.3 Extensions and Specification Curves

The same format of nested loops were built to iterate over every combination of control variables as well as the extra variables described in section 3.4.1. The coefficients were then extracted before the logarithm was taken to arrive at the odds ratio. This resulted in 7,200 models in each of the 6 specification curves for each odds ratio derived, 3 each for boys and girls. These are presented from figures 3.10 to 3.12 for boys, and from 3.13 to 3.15 for girls. The blue line show the magnitude of the odds ratios computed, and the light blue area show the 95% confidence intervals. The red dots represent odds in unadjusted models while the brown dots are for adjusted ones.

Figure 3.10: Specification Curve of Two Waves Odds Ratios (Boys)

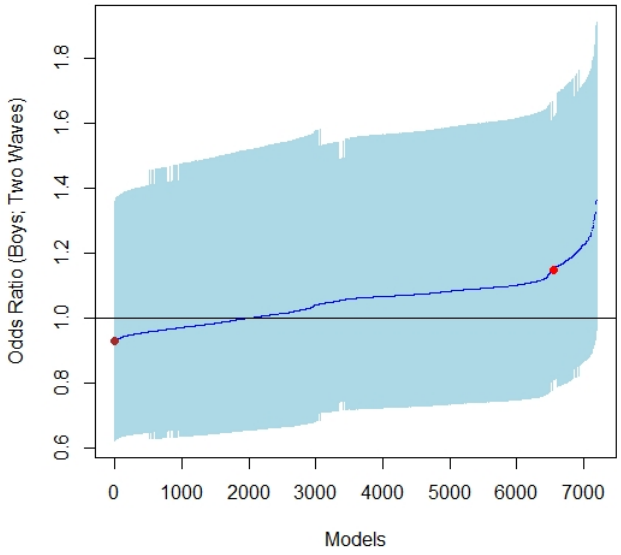


Figure 3.11: Specification Curve of One Wave Odds Ratios (Boys)

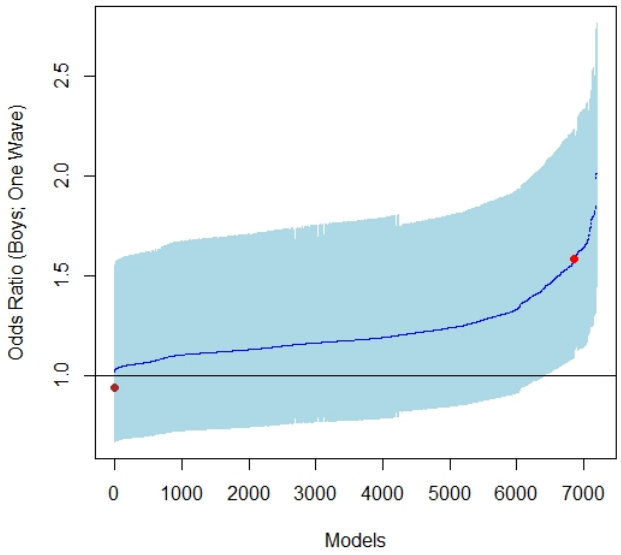


Figure 3.12: Specification Curve of No Waves Odds Ratios (Boys)

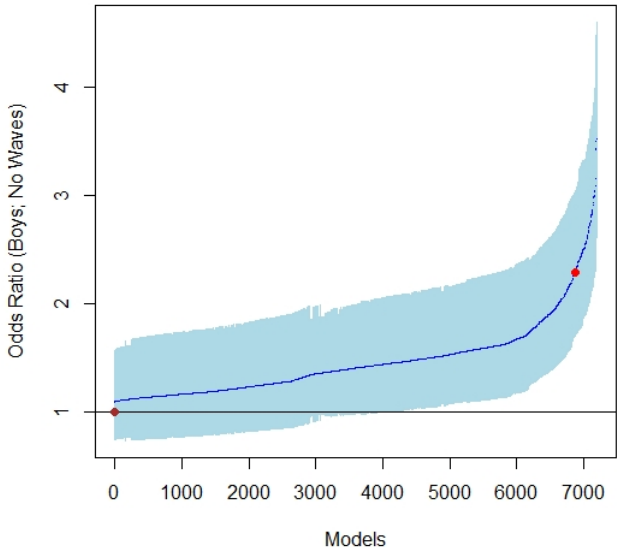


Figure 3.13: Specification Curve of Two Waves Odds Ratios (Girls)

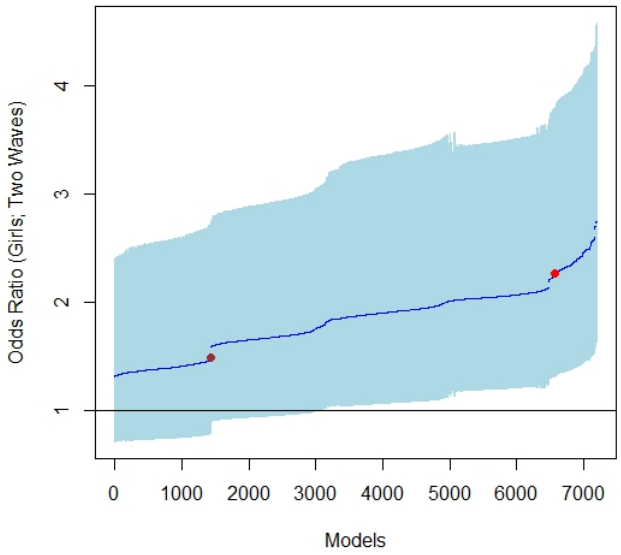


Figure 3.14: Specification Curve of One Wave Odds Ratios (Girls)

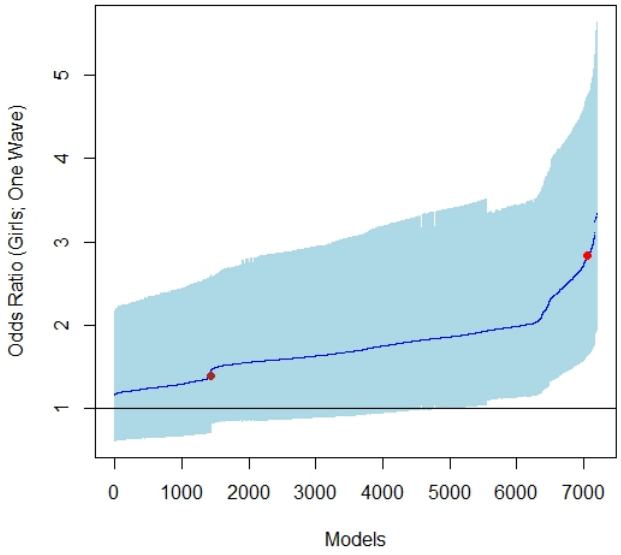
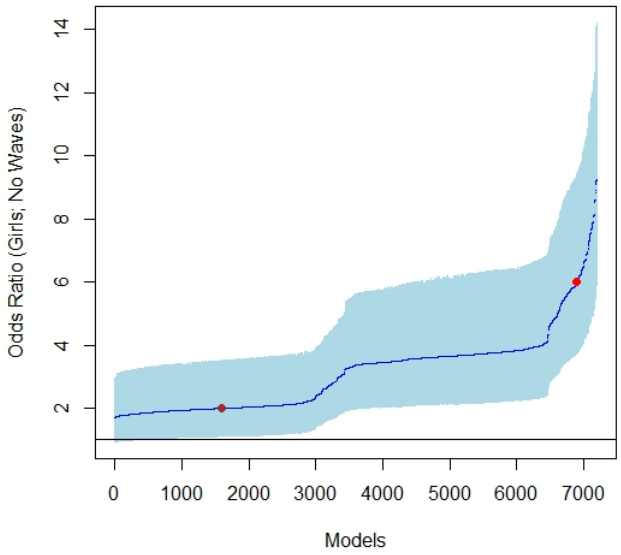


Figure 3.15: Specification Curve of No Waves Odds Ratios (Girls)



All odds ratios have upper tails which heavily skew upwards. This was more obvious the fewer waves that mothers were in paid work, and was more pronounced for girls than for boys. Apart from the tails (which represent models with either no or few controls), the specification curves were relatively flat. Furthermore, the inclusion of extra variables not used in the original study did not seem to be able to push the odds ratios far above or below the original study's findings. This indicates that the original study's odds ratios are stable and are not too sensitive to changes in control variables used.

For boys, the odds ratio of children whose mothers worked for two waves compared to three waves was between 0.92 to 1.36, which shows ambiguity in the direction of association when different controls are used. The direction is less ambiguous for the odds for one wave and no waves, as all 7,200 odds in the specification curve are above 1 (even if there is considerable overlap for the 95% confidence intervals).

The specification curves for girls were consistently of a larger magnitude than for boys across all maternal employment wave counts. The curves show that the odds of girls having socio-emotional behavioural problems are more sensitive to change with respect to inclusion and choice of control variables. While there were no abrupt breaks or sudden spikes in the curves for boys, this was present in all three curves for girls. For all 6 ORs, the original study's odds were spread out across the length of the specification curve.

This replication judges this study to have been successfully replicated and its odds ratios to be robust with respect to the use of control variables.

4 | Discussion

In the context of research on child socio-emotional behaviour, this paper highlights how sensitive some findings can be to seemingly arbitrary and atheoretical choices when constructing and choosing control variables. The strength of the association between different predictor variables can change substantially when researcher degrees of freedom are exploited. While a number of fields have had problems of unreplicable findings (Stevens, 2017), the specific sub-field around the determinants of child socio-emotional behaviour seem to be relatively replicable with three out of the four studies' headline results successfully reproduced. This may, however, boil down to the simplicity of the models used and that data was publicly available in a processed/cleaned form. This eliminates variability and further researcher degrees of freedom in how the data could have been pre-processed/cleaned or how complex models could be differently constructed.

Each of the four studies were replicated to different levels of success. Study 3 by Noonan et al. was the most successful because the authors were most transparent in presenting their data. By contrast, Study 1 by Rajyaguru et al. was the least transparent of the four and this was the study which could not be replicated. The more transparent the original study, the more replicable the results seem to be. Not only was Noonan et al.'s study the only one of the four with complete sample counts and descriptive statistics (Zilanawala et al.'s study had sample counts and rates of missingness only in relation to data imputation), but they had clearly described how every single variable was constructed from the original raw data. One thing which all studies could have done to make it even more transparent was to include the specific variable names that they had chosen (as Zilanawala did through private correspondence for some variables). Variable names used in this replication are available in the online R script.

The findings of this paper show that child socio-emotional behaviour has robust associations factors like long-term household income, maternal employ-

ment, and ethnicity. One surprising replicated finding was that compared to white British children, children of Black African ethnicity were likelier to have a lower SDQ TDS score, representing a tendency for fewer socio-emotional problems. This is despite literature on ethnicities usually portraying worse outcomes for ethnic minority children (Mosley & Thomson, 1995) – as was the case for other ethnicities studied like Indian, Pakistani, Bangladeshi, and Black Caribbean children.

Parental discipline strategies seem to have the weakest association with child socio-emotional behaviour. The original study claimed that, without conditioning on any control variables, a one unit increase in Active and Withdrawal discipline methods corresponded to a respective 1.11 and 0.72 unit rise in the socio-emotional behaviour problem score (the SDQ TDS). The replication found these figures to be 0.31 and 0.11 respectively. When controls are added, the coefficient sizes are even further reduced and are also far less statistically significant. The specification curves for Rajyaguru et al.'s study also had some sudden discontinuities which indicate the sensitivity of their coefficients.

Maternal employment and household income were clearly negatively associated with child socio-emotional behavioural problems. The longer that mothers were unemployed, the higher the odds that her child would have socio-emotional problems. Similarly, a higher permanent family income was associated with a lower probability of a child having socio-emotional difficulties. These findings were consistent with the results of the respective original studies, and their relatively flat specification curves (despite introducing new control variables in McMunn et al.'s study and Noonan et al.'s study having the most exhaustive list of control variables) show that the economic dimensions to child socio-emotional difficulties are the most robust ones.

5 | Conclusions and Limitations

Social science research that is transparent is reproducible. The more information the original studies had on how their variables were formed and chosen, the more replicable their final results were. Replicable studies are important because they bolster the strength of the claim(s) made in them.

One of the limitations of this analysis was that it did not model the changing statistical significance of different models as different sets of controls were used. Here, the only numbers that were extracted to build the specification curves were the coefficient sizes and the standard errors. One possible extension could be to observe how statistical significance varied as well by extracting either the t-statistics or the p-values. Statistical power could have also been modelled as sample sizes vary when different controls were used.

Another possible extension would be to include the different controls and key predictors each study used onto one another's models. This might have enabled the associations of each of their key explanatory variable to have been better compared. However, issues of over-fitting notwithstanding, this study did not use this approach as the SDQ TDS outcome used in different studies came from different waves which meant that it was not the same outcome variable used across the studies, thus making a direct cross-application unfeasible.

Bibliography

Christensen, A., & Heavey, C. (1987). Ainsworth, mds, blehar, mc, waters, e., & wall, s., patterns of attachment: A psychological study of the strange situation, hillsdale, nj: Erlbaum, 1978. ainsworth, mds, & eichberg, c., effects on infant-mother attachment of motherâs unresolved loss of an attachment figure, or other traumatic experience. in cm parkes, j. stevenson-hinde, & p. marris (eds.), attachment across the life cycle, london, routledge, 1991. *Psychiatry*, *144*(1), 1–9.

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. University of California Press.

Fernihough, A. (2011). Simple logit and probit marginal effects in r.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no âfishing expeditionâ or âp-hackingâ and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

Goodman, R. (1997). The strengths and difficulties questionnaire: a research note. *Journal of child psychology and psychiatry*, *38*(5), 581–586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337–1345.

Hansen, K. (2014). Millennium cohort study: a guide to the datasets. *First, Second, Third, Fourth and Fifth Surveys. London: Centre for Longitudinal Studies*.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L., ... Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine*, *32*(6), 959–976.

Lundborg, P., Nilsson, A., & Rooth, D.-O. (2014). Adolescent health and adult labor market outcomes. *Journal of Health Economics*, *37*, 25–40.

McMunn, A., Kelly, Y., Cable, N., & Bartley, M. (2012). Maternal employment and child socio-emotional behaviour in the uk: longitudinal evidence from the uk millennium cohort study. *J Epidemiol Community Health*, *66*(7), e19–e19.

Mosley, J., & Thomson, E. (1995). Fathering behavior and child outcomes: The role of race and poverty.

Noonan, K., Burns, R., & Violato, M. (2018). Family income, maternal psychological distress and child socio-emotional behaviour: Longitudinal findings from the uk millennium cohort study. *SSM-population health*, *4*, 280–290.

Platt, L. (2012). How do children of mixed partnerships fare in the united kingdom? understanding the implications for children of parental ethnic homogamy and heterogamy. *The Annals of the American Academy of Political and Social Science*, *643*(1), 239–266.

Rajyaguru, P., Moran, P., Cordero, M., & Pearson, R. (2019). Disciplinary parenting practice and child mental health: Evidence from the uk millennium cohort study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(1), 108–116.

Rutter, M., Tizard, J., & Whitmore, K. (1970). *Education, health and behaviour*. Longman Publishing Group.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *Available at SSRN 2694998*.

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in psychology*, *8*, 862.

Straus, M. A. (2017). Measuring intrafamily conflict and violence: The conflict tactics (ct) scales. In *Physical violence in american families* (pp. 29–48). Routledge.

Zilanawala, A., Sacker, A., Nazroo, J., & Kelly, Y. (2015). Ethnic differences in children's socioemotional difficulties: findings from the millennium cohort study. *Social Science & Medicine*, *134*, 95–106.

A | Strengths and Difficulties Questionnaire

This is the list of 25 questions asked in the SDQ. Questions ask caregivers (usually the mother) the extent to which their child displays a given behaviour or trait with three possible responses: Not True, Somewhat True, and Certainly True.

The first four categories of items are summed to create the Total Difficulties Score (TDS). Italicised items are reverse scored when summing.

Hyperactivity

- Restless, overactive, cannot stay still for long
- Constantly fidgeting or squirming
- *Thinks things out before acting*
- *Sees tasks through to the end, good attention span*
- Easily distracted, concentration wanders

Conduct Problems

- Often has temper tantrums or hot tempers
- *Generally obedient, does what adults request*
- Often fights with other children or bullies them
- Can be spiteful to others
- Often argumentative with adults

Emotional symptoms

- Often complains of headaches, stomach aches, or sickness
- Many worries, often seems worried
- Often unhappy, down-hearted or tearful

- Nervous or clingy in new situations, easily loses confidence
- Many fears, easily scared

Peer relationships

- Rather solitary, tends to play alone
- *Has at least one good friend*
- *Generally liked by other children*
- Picked on or bullied by other children
- Gets on better with adults than with other children

Prosocial behaviour (not included in TDS score)

- Considerate of other people's feelings
- Shares readily with other children (treats, toys, pencils, etc.)
- Helpful if someone is hurt, upset or feeling ill
- Kind to younger children
- Often volunteers to help others (parents, teachers, other children)