# Essays in family and labour economics

Yiyang Luo

A thesis submitted for the Degree of

*Doctor of Philosophy*

Department of Economics

University of Essex

February 2017

# Contents

# List of Figures

# List of Tables

*"Why worry*

*There should be laughter after pain*

*There should be sunshine after rain*

*These things have always been the same*

*So why worry now*

*Why worry now.*

*"*

*Dire Straits "Why worry"*

# Acknowledgements

Firstly, I would like to express my sincere gratitude to Professor Macro Francesconi, my advisor, for his continuous support of my Ph.D study, for his patience, intellectual input and excellent supervision. His guidance and encouragement were important and helpful in all the time of my doctoral research and job market. I am also very grateful to my thesis examiners — Professor Emilia Del Bono and Professor Kjell Salvanes. They have made valuable comments on the thesis and I have had useful discussion with them. Similarly, profound gratitude goes to other lecturers at Essex for sharing their expertises so willingly, in particular, Joao Santos Silva, Matthias Parrey and Thomas Crossley.

The provision of University of Essex Social Sciences Doctoral Scholarship is deeply appreciated, without this support, the completion of this thesis would not have been possible. I also thank the Department of Economics for providing funds for conference and training, which contributes to the improvement of this work.

I thank my fellow Ph.D. student, Aline, Daniel and Giulia, for their constant research and emotional support, for the stimulating discussions, for the days and nights working together before deadlines, for all the fun and stress we have had in the last four years. Also I wish to thank Eric, Simyee, Kevin, Ricey and the rest of the badminton team, with whom we spent lovely weekends together for the past two years. I thank Huajing, Kehai, Xiaoyu, Yang, Lin, Can, Bagel and Jianjun, with whom I felt the home culture in England. In particular, I am grateful to Rui and Yujun for sharing the ups and downs of the life abroad, regardless of the time difference and spatial distance. I also thank other numerous friends who played an essential part of my life at Essex.

I am hugely indebted to my family who showed unconditional support during my study away from home. Although they understood quite little about what exactly the Ph.D. life entails, they have always kept the faith in my choices. I apologise for the time lost with my retired mother and father.

# Summary

This thesis explores family and labour economics issues in the context of different countries, the unified motive is to gain policy implication by applying diversified micro-econometric tools into different datasets.

The UK has experienced the 1999 Working Family Tax Credit and the 2003 Working and Child Tax Credit reforms. The first chapter provides the first piece of evidence on the effect of single mothers being eligible to income transfer programmes on early childhood outcomes in the Britain. Using the Millennium Cohort Study (MCS), various children's production functions are used to deal with endogeneity of inputs and unobserved heterogeneity problems. Findings suggest that mothers entitled to in-work benefit has positive effects on both children's cognitive and non-cognitive outcomes, comparing to the mothers live on welfare.

The second chapter presents new evidence on the child quantity-quality (Q-Q) trade-off based on the 1% sample of 1990 Chinese census. The main contribution of this chapter comes from applying a novel Generalised Method of Moments (GMM) approach that accounts for the non-linear distribution of both outcome and endogenous variables. The identification strategy exploits variation in family size that is induced by twin births and first child gender, which allows the test of Q-Q trade-off in a wide range of fertility distribution. I find significantly negative effects of fertility on educational outcome of children, and this trade-off nonlinearly decreases with family size and shows heterogeneous effects by birth order. This chapter provides technique foundation for policies that attempt to reduce contraceptive costs, control population growth and subsidize families with fewer children.

The third chapter examines the retirement consumption puzzle using the Chinese Household Income Project data. A failure to smooth the consumption upon retirement would arise considerable concerns for the

well-being of elderly people and adjustments of public policies. This chapter employs a regression discontinuity approach and shows that elderly households are able to maintain stable consumption onset of retirement by adjusting expenditure across sub-aggregated categories and household behaviour. This study confirms the prediction of Life Cycle Model and have important implications for using disaggregated consumption data to test the existence of retirement consumption puzzle and for testing consumption theories.

# Chapter 1

## On Welfare or In Work: Perspective from Single Motherhood and Early Childhood Outcomes


### Abstract

This paper provides the first piece of evidence on the effect of being eligible to receive in-work benefit (WFTC and WTC/CTC) on the early development of children using single mothers from the UK Millennium Cohort Study (MCS). The treatment and control groups are defined based on the tax credit eligibility according to benefit rules. Various children's production functions are used to deal with endogeneity of inputs and unobserved heterogeneity problems. Findings suggest that mothers entitled to in-work benefit has positive effects on both children's cognitive and non-cognitive outcomes, comparing to the mothers live on welfare. Possible mechanisms through which the outcome has been improved are provided. In particular, an increase in formal childcare that is induced by entitlement to in-work benefit increases children's cognitive scores, and an increase in maternal time inputs improves children's non-cognitive skill. However, this paper fails to document any eligibility induced income effect. Additionally, the positive impact of being eligible to in-work benefit on non-cognitive score only exists if mothers work no more than 30 hours, which casts doubts on the 30 hour element of the in-work benefit policies. This study also suggests that the effect of being eligible to 2003 WTC/CTC are consistent with the 1999 WFTC, under a difference-in-difference framework.



**Key Words**: in-work benefit reform, tax credit, eligibility, child production function, cognitive outcome, non-cognitive well-being
**JEL Classification:**: C23, C52, H31, I24, J13

## 1.1 Introduction

Some developed countries—including the US, UK, Germany and Canada—have experienced origins and expansion of in-work benefit reforms in last twenty years. These reforms tend to increase the labour market participation of low-middle income households who rely on welfare and to reduce the inequality among children. The effects of those reforms have been widely analysed, especially in the sociology and economics literatures. Previous research has focused on the impacts on labour market outcomes and behavioural responses of the socio-economically disadvantaged individuals.

The UK has experienced two in-work benefit reforms in the last 20 years. First, the Family Credit (FC) was replaced by the Work Families' Tax Credit (WFTC) in October 1999. It has been concluded that WFTC gives more incentive for labour market participation than FC in various aspects (Brewer et al., 2009). In April 2003, the Work Tax Credit (WTC) and Child Tax Credit (CTC) have been introduced and replaced WFTC(figure 1.1). This paper regards 2003 WTC/CTC reform as a pure successor of 1999 WFTC for the following reasons. First, this paper focuses on the in-work incentive element of the reform, i.e. the sample considers lone mothers who receive WFTC and who receive both Working Tax Credit and Child Tax Credit later on as eligible, thus the differences between WFTC and WTC/CTC benefit rules are not expected to affect the sample. Second, this paper employs difference-in-difference approach and finds that the effects of being eligible to 2003 WTC/CTC on children's outcome is consistent with the effect under WFTC regime (robustness check session).

While researchers have explored wide questions that were raised by the in-work benefit reforms, there is very limited knowledge about the intergenerational consequences of those welfare reforms. Economists have recently contributed a few papers (Francesconi et al., 2008; Gregg et al., 2009) to addressing this issue in the UK. Unlike the impact on employment, the expected effects on next generation responses are ambiguous. On one hand, increasing parental investment driven by increased employment rates and higher income would have a positive effect on child's outcome. On the other hand, decreased parent-child interactions and parental supervisions resulting from spending more time working would have a negative effect on child's outcome. Additionally, increasing childcare usage induced by labour supply adjustment has ambiguous impacts on child outcome. Both theoretical predictions and empirical work provide evidence that family economic conditions at age 5 are more strongly associate with their adolescence education achievement than at child age 6 to 15 (Duncan et al., 1998; Shonkoff et al., 2000). Heckman (2006) also provides evidence on the impacts of early inputs on the achievement of child, adolescent, and adult. Although life cycle skill formation is a dynamic process, early environments strongly affect the productivity of later inputs. Thus changes in parental economic and employment circumstances at the early stage development of children are likely to affect their early and later on outcomes.

Since there has been far less analysis of the effect of these reforms on children's outcomes, this study aims at filling in this gap by providing the first piece of evidence on the effects of being eligible to the 1999 WFTC and later on 2003 WTC/CTC policies on children living in lone parent households. In particular, this paper focuses on young children's cognitive achievements and emotional outcomes that are critical to children's development and are also key predictors of later life achievements.

There are four reasons that this paper focuses on children of lone parents. First, there is an increasing trend of lone parenthood families in the UK. Lone parents with dependent children accounted for about 22% of all families with dependent children in 1996, increased to 25% in 2005 and 2015. Women represented 90% of lone parents with dependent children in 1996 and this percentage has barely changed since then[1]. Second, lone mothers are among the main target of in-work benefit policies, and literature has shown that the employment rate of lone mothers were increased by around 5 percentage points from 1993 to 2004 and working hours among those already in work have also increased (Gregg et al., 2009). Thus low-paid single mothers are most likely to be lifted out of poverty by tax credit policies supplementing their wages. This group is the most relevant for studying whether the tax credit reform reduces welfare dependency (Eissa and Liebman, 1996). Third, there is less to worry about the joint labour supply decision of other family members. There is evidence that the probability of participation is reduced for the secondary earner through income effects in a two-parent family. Fourth, children in lone parents families have lower extent of happiness, self-esteem, lower quality relationships with the mother and worse behaviour problem (Gregg et al., 2009).

It is useful to understand whether those children of single mothers who benefit from these reforms have better outcomes than those of welfare families in the early childhood perspective. Since children in the sample were born in September 2000-2001, they were exposed to the implementation effects of the WFTC reform at 9 months and age 3[2], and affected by WTC/CTC at 5 and 7 years old. Therefore, lone parent

[1]This information is obtained from Families and Households 2015 report of Office for National Statistics(ONS).

[2]Due to different interview date, the information documented at wave 2(age 3) could refers to the outcome at before/after the implementation of 2003 WTC/CTC.

households are the relatively vulnerable group in the society and also the main target group of these reforms. The children growing up in this environment are likely to be affected in a wide range of aspects. However, due to the highly under reported number of actual recipients of in-work benefits, moreover, due to the fact that no available dataset in Britain contains both comprehensive measurement of child outcomes and good quality family background information, before and after the 1999 WFTC reform, this paper identifies the effect of being eligible to in-work benefit on children's outcomes, irrespective of take up. It captures the differences on the development of children among three groups of households. The treatment group is low-income in-work group, which is defined according to the benefit eligibility, i.e. lone mothers who work at least 16 hours per week and have income below the eligibility threshold. This paper considers two control groups. The first control group is low-income on-welfare group, which consists of those mothers who are not in the labour market or work less than 16 hours per week. It is ineligible to in-work benefit because mothers must work no less than 16 hours in order to get WFTC tax benefit. Under the WTC/CTC reform, CTC is also available to families without requirement on labour market participation. Since this paper is interested in in-work benefit, mothers who are only eligible to CTC are included in this control group. The second control group is the middle-income in-work group. Mothers of this group work no less than 16 hours but are ineligible for receiving tax credit due to their income is beyond the income threshold. This paper does not consider high income households since they are likely to have considerable different inputs in children. Details on the eligibility rule are provided in the section 1.2.1 policy background.

By comparing the children of three groups of mothers, it is informative on whether mother is eligible to receive in-work benefit improves next generation's well-being to the extent that income, working hours, maternal time input and childcare usage affect child achievement. If the cut-off of benefit eligibility is not fully overlapping with observed and unobserved differences among groups, a positive effect found on the treatment group, with the first control group as the reference group, would suggest a positive association between the tax benefit and children's outcome. Moreover, under the assumption that the observed and unobserved differences of characteristics are fixed over time, it is possible to attribute any effects found to the treatment effects under a child fixed effect framework.

This paper contributes to the literature in several dimensions. First, it provides the first piece of evidence on the effects of being entitled to in-work benefit—1999

WFTC and 2003 WTC/CTC—on early childhood outcomes for single mothers using longitudinal data drawn from the Millennium Cohort Study (MCS). The impact of those two policies on very young children has not been examined, to the best of my knowledge. By using longitudinal structure, this study is able to control for changing sample composition over time, which is resulting from changes in tax credit rules. Results suggest that entitlement to in-work benefits improve children's cognitive outcomes by 0.078 and non-cognitive well-being by 0.071 of one standard deviation, comparing to children of on-welfare mothers. Second, this paper also explores several mechanisms through which the outcome has been improved. In particular, an increase in formal childcare that is induced by entitlement to in-work benefit increases children's cognitive score by 0.132 of a standard deviation, and an increase in maternal time inputs improves children's non-cognitive skill by 0.051. Third, the results suggest no differential improvement of children's outcome that is due to different location at the income distribution of treatment group. It is in contrast with the finding of Miller and Zhang (2009) and Dahl and Lochner (2012) that the income effects are larger for children from poorer family. Forth, further investigation shows that the positive impact of being eligible to in-work benefit on non-cognitive score only exists if mothers work no more than 30 hours, which casts potential concern on the 30 hour element of the in-work benefit policies. Fifth, this study also suggests that the effect of being eligible to 2003 WTC/CTC are consistent with the 1999 WFTC, under a difference-in-difference framework. It is probably due to the fact that the changes in policy parameters have negligible effect on working status of lone mothers.

The remainder of this paper is organized as follows. Section 1.2 presents the literature review on the analysis of welfare reform effects. Section 1.3 states the theoretical framework that this paper focuses on. Section 1.4 describes the data source and descriptive statistics. Section 1.5 provides main results and robustness checks. Finally, Section 1.6 draws conclusion.

## 1.2 Policy Background and Literature Review

### 1.2.1 Policy Background

The UK government has launched two in-work benefit reforms to try to help individuals with low income in the last 20 years. A timeline of these two reforms and sample observation periods are indicated in figure 1.1. The key objectives of those in-work benefits are to increase in-work incentives of low-income groups and reduce child poverty. First, the Family Credit (FC) was replaced by the Working Families' Tax Credit (WFTC) in October 1999. Prior to 1999, FC is the benefit that supports low-income working families. The successor WFTC is referred to as tax credit rather than a benefit, and the eligibility depends on working hours and existence of dependent child(ren) in the household. Specifically, families with dependent children[1] and with at least one adult work more or equal to 16 hours per week are entitled to claim tax credit[2]. The amount of credit varies with number of dependent children, earnings, working hours and savings. The first income(net) threshold was initially £90 per week with taper rate 55%, an extra credit can be earned if at least one adult worked at least 30 hours per week.

WFTC was more generous than its predecessor in four aspects (Adam and Brewer, 2005; Brewer et al., 2009). First, it provided higher credits, especially for those younger children age 0-10. Second, the net income threshold was increased substantially. Third, the taper rate (tax rate on earnings) at which the maximum credit was reduced is lower. Fourth, it provided a large separate childcare tax credit. Under FC, the childcare costs, up to £60(£100) per week could be disregarded from the calculation of total net income for families with 1 child(2 children). WFTC offered a separate childcare tax credit rather than an income disregard, and it covered 70% of childcare costs up to £100(£150) per week for 1 child(2 children).

In April 2003, the Work Tax Credit (WTC) and Child Tax Credit (CTC) have been introduced and replaced its predecessor WFTC. WTC differs from WFTC in that the former provides in-work support to families without children as well. CTC is not an in-work benefit, i.e. it is a means-tested benefit and provides childcare benefit to families with no members working as well. Since this paper fails to document any impact of the transitioning from WFTC to WTC/CTC benefit on the employment rates of lone mothers, this paper assumes that these two reforms are coherent and can be regarded as post FC in-work benefits in general. Another important difference between and WFTC and WTC/CTC is the change in the definition of recipient. The claimant received the credit under WFTC, while the childcare element of WTC and CTC are made to the main carer of the child(ren). The main carer is identified by the couple themselves. Similarly, the WTC credit is paid to the main earner who works more than 16 hours

[1]In the UK, a dependent child is defined as younger than age 16 or younger than age 19 if in full-time non-university education.

[2]The eligibility rule also depends on family savings and capital.

per week. The couple must decide who is the main earner if both work 16 hours or more. Her Majesty's Revenue and Customs (HMRC) decides mother as the main carer and father as the main earner if couples do not decide by themselves or cannot reach an agreement. Since my subject of interests are children of lone parents, this difference should be minor. Thus this paper will not distinguish the effects of being eligible to receive either WFTC or WTC/CTC.

For ease of understanding the parameters of the reform, the tax credit entitlement can be calculated in the following formula:

$$Credit = A - max(Y - \vee, 0) * T \tag{1.1}$$

where $Credit$ is the amount of entitlement, $A$ is the maximum amount of tax credit a family can be awarded, $Y$ is family net income, $\vee$ is the income threshold above which tax credit would be withdrawn, at taper rate $T$.

This paper defines treatment and control group depending on the sample eligibility of tax credit entitlement, as shown in equation 1.1. As tax credit is decreasing in net income, a level of income can be obtained at which entitlement reaches zero for the average family in the sample (Fisher, 2016). Under the WFTC regime, tax credit was withdrawn at a rate T=55% from net earnings, an average lone mother family is characterized as 1.78 children, £13.67 weekly childcare costs, tax credit reaches zero for a weekly net income of £314.79[1], according to table 1.1. Under the WTC/CTC reform, tax credit has a taper rate of 33% from gross earning, an average family with 2.02 children, £18.01 weekly childcare costs, WTC/CTC credit reaches zero for a weekly gross income of £400.14 and CTC credit reaches zero for a weekly gross income of £400.03[2], based on table 1.2. Due to the slight change in benefit rules, the treatment and control groups are varying across time.

Note that the actual treatment is made based on family earnings, number and age of dependent children, childcare costs, level of savings and hours of work, thus the actual entitlement is family-specific. There could be cases that families are assigned to control group but in fact they are eligible to claim tax credit, and cases that treated group are in fact receive no tax credit. One way to minimize the contamination effect is to exclude those observations at the eligibility margin.

[1] Apply 2001-2002 WFTC rule.
[2] Apply 2003-2004 WTC/CTC rule.

### 1.2.2 Literature Review

There are variety of studies exploring the effect of welfare-to-work policies on individual's outcomes, theoretically and empirically. Some studies find that the responses on labour market participation are as expected, while some report unintended or negligible consequences in partnership and child well-being. For the US, empirical work has focused on various aspects—especially labour market behaviour—of the in-work benefit and work-conditioned programmes (Grogger et al., 2009; Grogger and Karoly, 2009). While for the UK, limited evidence about other behaviour responses beyond employment changes has been provided (Brewer et al., 2009).

One of the key objectives of all the tax-benefit policies or programmes mentioned above is to improve work incentive and increase employment rate of low-income families, especially single mothers. Both structural model and difference-in-difference model (Blundell et al., 2005; Francesconi and Van der Klaauw, 2007; Gregg et al., 2009) show an increase in the employment rate of lone parent, at about a 4-7 percentage points, which was driven by both high rates at which they entered in the labour market and high rates at which they remained in it. WFTC also led to a significant increase in lone mothers' hours of work, which makes them become eligible for the benefit. Blundell et al. (2008) also point out that job changes contribute mostly to the labour supply adjustment. These evidence on labour market outcomes draw researchers' attention to the perspective of next generation.

For the US, a number of previous studies use the National Longitudinal Survey of Youth (NLSY) data to evaluate the policy impact on children's cognitive outcome, but fail to reach a consensus on the estimated results (Bernal, 2008). In fact, most of them fail to control for the endogeneity problem that some unobserved characteristic of mothers affect their employment and affect children's cognitive ability as well. Past research has found the impacts of in-work benefit and work-conditioned programmes vary by children's development periods. Duncan et al. (2000) and Zaslow et al. (2002) find favourable effects on school-age children in terms of cognitive and academic outcomes, through improvement of family economic status, while unfavourable effects on adults in terms of behavioural problem. In contrast, programmes without effects on parental income have few effects on school-age children (Clark-Kauffman et al., 2003). However, their results simply based on an OLS regression. The unobserved characteristics of mother and child that correlated with employment and welfare reforms might also affect changes in child outcomes.

Bernal (2008) introduces a structural model of single mothers' decisions about work and child care use, and how these decisions affect child cognitive outcomes. In general, single mother's utility is derived from consumption, employment, child's outcome and welfare participation. Results imply negative and sizeable effects of maternal employment and child care on cognitive outcomes. An additional year of full-time work and child care involves a 1.8 percent reduction in cognitive test score. And this effect is greater on high ability child in the sense that spending time with high ability children has higher technological return than time spent with low ability children. Building on this work, Bernal and Keane (2010) construct a quasi-structural model to take into account both the exogenous variation in employment and child-care decisions of mothers induced by the variation in welfare rules and child's production function. They find a 2.7% reduction (0.14 standard deviation) on child's cognitive ability score if the mother works full-time and uses full year child care. Similarly, but based on a reduced form framework, Bernal and Keane (2011) deal with endogeneity problem by constructing a set of welfare policy variables and using local demand conditions. Basically, they exploit the exogenous variation in income and childcare time that is derived from welfare policy rules. They find that child cognitive scores decrease 2.1% by a year of child care, and that formal child care has no adverse effect, but informal care has significant negative effects.

Grogger and Karoly (2009) provide a theoretical framework to explore the effects of work-conditioned transfer programmes on child well-being. Their idea, essentially, is to maximize single mother's utility subject to a budget constraint and a production function which associate child 'quality' to parental time and goods purchase (Becker and Tomes, 1976). Mother decides the time/income allocation between children and work/consumption goods. They also find that the effects vary with the age and outcome of interest of the children, due to the trade-off between increased family income and reduced parental supervision.

Miller and Zhang (2009) estimate the long-run effects of Earned Income Tax Credit (EITC) on the academic performance of children in low-income families. Their identification comes from comparing low-income and middle-income groups, where the former consists of students who are eligible for federal free lunches, and the latter group are not eligible. Based on a difference-in-difference method, they find the improvement rate of maths score of low-income family children is bigger than that of middle-income family. The same results have been found in their value-added framework.

Dahl and Lochner (2012) identify the effect of family income on child achievement based on the large, non-linear and exogenous changes in the Earned Income Tax Credit (EITC). They find that child test scores increase 6% of a standard deviation by raising 1,000 dollar in income from the contemporaneous model. Dynamic models find small effects from past period income while contemporaneous income has the largest effect on achievement. By estimating separate regressions for various subgroups, results suggest that the income effects are larger for children living in more disadvantaged households, younger children, and boys.

In Canada, the study of policy effect on children is mainly based on the National Longitudinal Survey of Children and Youth (NLSCY). Dooley and Stewart (2004) use OLS and fixed effects models to estimate the effects of family income on children's test scores. Their findings are similar to the result of Blau (1999) in that they find a small relationship between income and test scores. Based on an instrumental strategy, Milligan and Stabile (2011) examine the effects of family income on child well-being by exploiting changes in child benefits in Canada. In particular, they use the variation in child benefits across province, time, and family type. They find that child benefit programs have significant positive effects on test scores and mental health. Their finding also suggest that the policy effect varies with child gender: child benefit effects are bigger on mental health outcomes for girls, and on educational outcomes and physical health outcome for boys.

Results above give strong incentives to evaluate the impacts of being eligible to receive tax benefit on children's outcomes in the UK. Unfortunately, there have been far less analyses of the effect of tax-benefit reforms on early childhood outcomes, however, this is a crucial question from a policy point of view.

Francesconi et al. (2008) applies the bargaining household models (Chiappori, 1992) to characterize the noncooperative behaviour of divorced parents, and shows adverse effects of tax-benefit policies on those parents and their children's well-being. This is due to the fact that the in-work benefit policy increases mother's effective wage and reinforces the inefficiencies induced by noncooperation.

Using data from the British Household Panel Survey (BHPS) youth files, Gregg et al. (2009) include mother's mental health as an explanatory variable in the difference-in-difference regression. With the introduction of the WFTC, the policy effect on children is mixed. They find positive effects on boys, while no significant impacts on girls. In fact, WFTC has profoundly improved child well-being—self-esteem, unhappiness scores, smoking and planning to leave school at age 16 — in lone parent families through increased maternal employment and mental health. However, their estimation focus on adolescent children age from 11 to 15 years old.

Previous research provides estimation of welfare reform effects at various aspects, while this paper tries to extend existed knowledge about the effects of tax-benefit policies on child's cognitive and non-cognitive outcomes in the UK, with a particular focus on very young children.

## 1.3 Econometric Method

This section begins with introducing the identification strategy followed with various specification of child production function. The main interest of this paper is to understand the relationship between 1999 and 2003 in-work benefit rules and cognitive and non-cognitive outcomes at early stage of child development. On the one hand, one would expect child's outcome to be improved through increased family income(paid formal childcare use) or role model effects; on the other hand, the decline in maternal time inputs may severely damage child's early stage development and may have adverse consequences on child's achievement in the long run. The prediction and existing empirical finding on this relationship is ambiguous.

This paper defines treatment group as low-income in-work group, and the first control group as low-income on-welfare group and the second control group as middle-income in-work group. The second control group is more similar to the treatment group on one dimension—they have similar working hours—but less on another: they have much higher income(figure 1.2). Due to the lack of pre-reform data, it is difficult to separate income effects from policy changes and tease out any other contemporary policy changes. Additionally, this paper assumes full take-up rate due to the fact that individuals highly under-report actual benefit receipt and an exact calculation of the entitlement is beyond the scope of this paper's attempt. This paper aims to bridge the gap that there is far less evidence on the effect of in-work benefit eligibility on early childhood outcomes in the UK.

This paper compares the outcomes of children whose mothers were eligible for WFTC or later on WTC/CTC to other mothers who are ineligible, where the identification comes from differences in the intensity of treatment between the treatment and two control groups. Under this framework, the treatment and control groups are defined regardless of the treatment they actually received, and the subsequent deviation from eligibility or withdrawal from treatment[1] (Fisher et al., 1989). Arguably, the comparison among three groups provides some upper and lower bound estimation of the effect. The identification is derived in the following steps:

Build on Angrist and Pischke (2008), the observed differences in child outcome are given by

$$\begin{aligned} &E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \\ &E[Y_{1i} - Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \end{aligned} \tag{1.2}$$

Where $Y_{1i}$ represents child $i$'s outcome of lone mother families that have a high intensity of getting treated, $Y_{0i}$ means child $i$'s outcome of single mother families that have a low intensity of being treated. $D$ is an indicator for treatment. This assumption is correct if the treated and control groups had the same propensities for child development and faced the same environments, except for the presence of in-work benefits received by the treated group. The first part of the right hand side gives the average treatment effects on those get tax credit, and the second part provides a selection bias. If those who get tax credit would have better child outcome anyway, comparing to those who have low income but ineligible to tax credit(low-income on-welfare group), then the naive comparison $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ would exaggerates the benefits of the reform. Likewise, if those treated would have worse child outcome anyway, comparing to higher income ineligible working group(middle-income in-work group), then the simple comparison would underestimate the effects.

To control for intrinsic differences in group-specific compositional changes over time, a set of demographic controls are included in all the specifications, i.e. child gender, child birth weight, number of siblings, birth order, mother age at child birth and age square, education level of mother, ethnicity and region of residence. Under the conditional independence assumption, selection bias disappears conditional on observed characteristics $W_i$. Formally, this means

$$Y_{0i}, Y_{1i} \perp D_i | W_i \tag{1.3}$$

Given this assumption holds, the observed average differences in child outcome are

[1] This assumption will be relaxed in the robustness check session.

identified by

$$E(Y_i|W_i, D_i = 1) - E(Y_i|W_i, D_i = 0) = E[Y_{1i} - Y_{0i}|W_i] \tag{1.4}$$

The empirical framework that is adopted for the estimation of cognitive achievement production function conceptualizes child development as a cumulative process. Conventionally the test score gap is accounted for by including the historical or/and current family inputs, school inputs, parents' ability and child unobserved endowments. Todd and Wolpin (2007) show that family inputs are significant determinants of children's cognitive test scores. As it is suggested by Todd and Wolpin (2003, 2007), this paper includes maternal treatment status[1] and other characteristics that determine children's outcome. Let $Y_{it}$ denotes the observed test score measure for child $i$ at age $t$ and it is determined by group dummies $D_{it}$, a set of observed inputs $X_{it}$ up to age $t$ and unobserved inputs $Z_{it}$ up to age $t$, as well as children's ability endowment $\mu_i$. The main variables of interest are $D_{it}$ and its values at previous period. $\epsilon_{it}$ is the measurement error. Under the assumption that the production function is linear, the regression is specified as follows:

$$\begin{aligned} Y_{it} =& D_{it}\delta_1 + D_{it-1}\delta_2 + \cdots + D_{i1}\delta_t + X_{it}\alpha_1 + X_{it-1}\alpha_2 + \cdots + X_{i1}\alpha_t \\ & Z_{it}\beta_1 + Z_{it-1}\beta_2 + \cdots + Z_{i1}\beta_t + \delta_t\mu_i + \epsilon_{it} \end{aligned} \tag{1.5}$$

The empirical challenge for estimating equation 1.5 comes from three aspects. First, children's ability endowments are unobserved. Second, the full history inputs are likely to missing. Third, the measurement error in test scores. In order to address these problems, a variety specifications of equation 1.5 have been adopted in previous literature (Todd and Wolpin, 2003, 2007; Andrabi et al., 2009, 2011). In particular, Andrabi et al. (2009) try to correct measurement error problem by using the heteroskedastic standard error of each test score returned by Item Response Theory[2], and Andrabi et al. (2011) correct this problem by using alternate subjects as instruments. Both methods provide similar results.

[1]Childcare usage and maternal time inputs are not included at the main regression, as they might be potential outcomes of the in-work benefit eligibility. Those inputs are included later on as mechanisms through which eligibility to WFTC or WTC/CTC might affect children's development.

[2]Item Response Theory calculates standard error for each test score from an estimation based inverse information matrix.

### 1.3.1 Contemporaneous Specification

A bunch of early economics studies focus on current inputs by placing very strong restrictions on equation 1.5. Equation 1.6 gives the contemporaneous specification.

$$Y_{it} = D_{it}\gamma_1 + X_{it}\alpha_1 + e_{it} \tag{1.6}$$

The error term can be rearranged by combining equation 1.5 and equation 1.6,

$$\begin{aligned} &e_{it} = D_{it-1}\gamma_2 + \cdots + D_{i1}\gamma_t + X_{it-1}\alpha_2 + \cdots + X_{i1}\alpha_t \\ &Z_{it}\beta_1 + Z_{it-1}\beta_2 + \cdots + Z_{i1}\beta_t + \delta_t\mu_i + \epsilon_{it} \end{aligned} \tag{1.7}$$

Equation 1.7 indicates that the assumption required to consistently estimate $\gamma_1$ and $\alpha_1$ is that all the omitted factors including observed lagged inputs, unobserved current and past inputs, initial endowment, and measurement error are uncorrelated with included inputs $D_{it}$ and $X_{it}$.

### 1.3.2 Cumulative Specification

Based on the contemporaneous specification, there are two main specifications which are cumulative specification and value added specification. In general, all the other variations are derived from them.

Cumulative specification augments the contemporaneous model by including past observed inputs.

$$Y_{it} = D_{it}\gamma_1 + D_{it-1}\gamma_2 + \cdots + D_{i1}\gamma_t + X_{it}\alpha_1 + X_{it-1}\alpha_2 + \cdots + X_{i1}\alpha_t + e_{it} \tag{1.8}$$

The error term can be derived by combining equation 1.5 and equation 1.8.

$$e_{it} = Z_{it}\beta_1 + Z_{it-1}\beta_2 + \cdots + \delta_t\mu_i + \epsilon_{it} \tag{1.9}$$

It works under the assumption that any omitted inputs and initial endowments are uncorrelated with included inputs.

### 1.3.3 Value Added Specification

Value added specification augments the contemporaneous model by including lagged outcome, which can be viewed as a sufficient statistics for the omitted past inputs,including any past unobserved endowments. It is often adopted when data on past inputs are partly or completely missing.

$$Y_{it} = D_{it}\gamma_1 + X_{it}\alpha_1 + \theta Y_{it-1} + e_{it} \tag{1.10}$$

By multiplying lagged outcome $Y_{it-1}$ with $\theta$, I get

$$\begin{aligned} \theta Y_{it-1} =& \theta D_{it-1}\gamma_1 + \theta D_{it-2}\gamma_2 + \cdots + \theta D_{i1}\gamma_{t-1} + \theta X_{it-1}\alpha_1 + \theta X_{it-2}\alpha_2 + \cdots + \\ & \theta X_{i1}\alpha_{t-1} + \theta Z_{it-1}\beta_1 + \theta Z_{it-2}\beta_2 + \cdots + \theta Z_{i1}\beta_{t-1} + \theta\mu_i\delta_{t-1} + \theta\epsilon_{it-1} \end{aligned} \tag{1.11}$$

Then subtracts equation 1.5 with $\theta Y_{it-1}$ in both sides,

$$\begin{aligned} Y_{it} - \theta Y_{it-1} =& \gamma_1 D_{it} + (\gamma_2 - \theta\gamma_1)D_{it-1} + \cdots + (\gamma_t - \theta\gamma_{t-1})D_{i1} + \\ & \alpha_1 X_{it} + (\alpha_2 - \theta\alpha_1)X_{it-1} + \cdots + (\alpha_t - \theta\alpha_{t-1})X_{i1} + \\ & \beta_1 Z_{it} + (\beta_2 - \theta\beta_1)Z_{it-1} + \cdots + (\beta_t - \theta\beta_{t-1})Z_{i1} + \\ & (\delta_t - \theta\delta_{t-1})\mu_i + (\epsilon_{it} - \theta\epsilon_{it-1}) \end{aligned} \tag{1.12}$$

The error term in equation 1.10 can be expressed as

$$\begin{aligned} e_{it} =& (\gamma_2 - \theta\gamma_1)D_{it-1} + \cdots + (\gamma_t - \theta\gamma_{t-1})D_{i1} + \\ & (\alpha_2 - \theta\alpha_1)X_{it-1} + \cdots + (\alpha_t - \theta\alpha_{t-1})X_{i1} + \\ & \beta_1 Z_{it} + (\beta_2 - \theta\beta_1)Z_{it-1} + \cdots + (\beta_t - \theta\beta_{t-1})Z_{i1} + \\ & (\delta_t - \theta\delta_{t-1})\mu_i + (\epsilon_{it} - \theta\epsilon_{it-1}) \end{aligned} \tag{1.13}$$

Equation 1.13 indicates that this specification also imposes a set of restrictions on estimating the production function. First, the coefficients on the past observed inputs and the coefficient on all the unobserved (current and past)inputs are age-invariant (i.e. $\gamma_t = \theta\gamma_{t-1}$, $\alpha_t = \theta\alpha_{t-1}$, $\beta_t = \theta\beta_{t-1}$ for all t). Second, the error term include child heterogeneity in learning, and we assume that the heterogeneity enters through a one-time process. According to $\delta_t = \theta\delta_{t-1}$, the coefficient associated with child endowment is constant scross age. Third, as suggested by Todd and Wolpin (2003, 2007), in order to consistently estimate $\gamma_i$ and $\alpha_i$, $\epsilon_{it}$ must be serial correlated in a way that makes $\epsilon_t - \theta\epsilon_{t-1}$ become an i.i.d shock. Otherwise, $Cov(Y_{it-1}, e_{it}) = 0$ is violated.

Since dynamic panel model does not address measurement error in test scores on its own, Andrabi et al. (2011) state that the measurement error in the value added specification will bias the estimated effect on inputs and lagged outcome. In particular, the measurement in the lagged test score will cause an attenuation bias. Letting $Y_{it} = Y_{it}^* + u_{it}$ denote the observed outcome, where $Y_{it}^*$ is the true outcome and $u_{it}$ is the measurement error in test scores. Equation 1.10 could be rewritten as

$$Y_{it}^* = D_{it}\gamma_1 + X_{it}\alpha_1 + \theta Y_{it-1}^* + e_{it} + u_{it} - \theta u_{it-1} \tag{1.14}$$

Therefore, the measurement error in the test score causes a downward bias specified as

$$plim\theta = \theta(1 - \frac{\sigma_u^2}{\sigma_{y^*}^2 + \sigma_u^2}) \tag{1.15}$$

Ideally, a valid instrument for lagged one period test score would be previous period test scores $Y_{is}(s <= t-3)$. Due to the restriction of the data availability, outcome $Y_{it-1}$ could be instrumented by alternative outcomes, assuming that they are subject to different sources of measurement error.

### 1.3.4 Cumulative Value Added Specification

Cumulative value added specification relaxes the assumption required by value added model in the sense that the cumulative value added specification does not need the age-invariant effects of the observed inputs. Cumulative value added specification also relaxes the assumption requires by cumulative specification in the sense that the value added cumulative specification allows the effects of past observed outcomes. It is reasonable that parents chooses current inputs based on past observed outcomes. Cumulative value added specification can be expressed as

$$Y_{it} = D_{it}\gamma_1 + D_{it-1}\gamma_2 + \cdots + D_{i1}\gamma_t + X_{it}\alpha_1 + X_{it-1}\alpha_2 + \cdots + X_{i1}\alpha_t + \theta Y_{it-1} + e_{it} \tag{1.16}$$

The error term is obtained by multiplying equation 1.5 with $\theta$ and then subtract equation

$$\begin{aligned} e_{it} =& \beta_1 Z_{it} + (\beta_2 - \theta\beta_1)Z_{it-1} + \cdots + (\beta_t - \theta\beta_{t-1})Z_{i1} + \\ & (\delta_t - \theta\delta_{t-1})\mu_i + (\epsilon_{it} - \theta\epsilon_{it-1}) \end{aligned} \tag{1.17}$$

The error term given by equation 1.17 indicates the restriction imposed in this specification is same as value added model except allowing for the effect of observed inputs to change across time.

## 1.4 Data Source

The data for this study is from the UK Millennium Cohort Study (MCS), which began as a longitudinal study of 18,819 babies in 18,533 households in the UK. All of these children were born between 1 September 2000 and 31 August 2001 in England and Wales, and from 23 November 2000 to 11 January 2002 in Scotland and Northern Ireland. The first sweep of the survey was conducted during 2001-2002 and collected from parents when the babies were about 9 months old. The sample design allowed for over-representation of individuals from ethnic minorities and families living in areas with high rates of child poverty. The families were followed up at approximately two-year interval. Subsequently, the second sweep was carried out when the children were aged about 3 years, and the third one was collected when they were aged about 5 years old. When children were 7 years old, the information was collected in the fourth sweep. For ease of understanding, figure1.1 indicates each sample collection period and the introduction of each in-work benefit policy. This dataset provides repeated measurement and assessment of children which allows me to account for unobserved child fixed effects, and it includes rich information on mother's and child's characteristics, sources and amount of family income, and various family inputs.

The sample consists of natural mothers who are the main respondent across the four interviews and who give birth to only one cohort child(multiple births are not considered). The sample is further restricted to mothers who are single (legally separated, single never married, divorced and widowed) during the 7 years' observation period[1] and at least 16 years old at child birth, and of whom the child has at least two cognitive or non-cognitive measures at age 3, 5 and 7. After pooling all 4 waves, the resulting sample for cognitive score and noncognitive score is 4775 and 4820 observations, separately.

As mentioned in section 1.1, this paper considers three groups of single mothers depending on tax credit eligibility, which are the low-income in-work group(treatment group), on-welfare group (the first control group) and middle-income in-work group(the second control group).

This paper uses the British Ability Scales (BAS) to measure children's cognitive

[1] This means the mother is reported to be single at interview across 4 waves.

outcomes. BAS have been administered to test cognitive abilities and educational performance for children aged from 2 years 6 months to 7 years 11 months, which are of great importance in helping us to understand the core aspects of child's ability: verbal, picture reasoning and spatial abilities. This study uses the BAS Naming Vocabulary Test to measure cognitive outcome at age 3 and 5, and uses BAS Word Reading at age 7. To provide comparability in outcomes across studies, this paper transforms the ability scores into Z-score, with mean 0 and standard deviation 1, which is adjusted for age and difficulty of the questions.

The MCS uses the Strength and Difficulties Questionnaire (SDQ) to measure children's behavioural development. The SDQ is a behavioural screening questionnaire for children aged from 3 to 16. It measures five sub-scales—emotion symptoms, conduct problems, hyperactivity, peer problems and pro-social behaviour—each of the sub-scale consists of 5 items. This study employs the Total Difficulty Score, which is obtained by summing up the first four sub-scales, to measure children's non-cognitive outcomes. The Total Difficulty Score ranges from 0 to 40 and higher scores represent more problematic behaviour. For the ease of interpretation, this paper reversely codes the SDQ score – higher non-cognitive score means less behavioural problems. To provide comparability in outcomes across studies, the Total Difficulty Score are standardized with mean 0 and standard deviation 1.

The control variables included in this paper are child age at interview date, gender, birth weight, number of siblings, birth parity, ethnicity, region of birth, mother's education level, age at child birth.

Mother's working hour and income level by group and child age are provided in figure 1.2. The treatment group has an income about 3000 pounds more per year than low-income on-welfare group, and this income gap is persistent across years. Treatment group earns annual income about 10000 pounds less than the middle-income in-work group(panel (a)). Panel (b) shows that the average working hours for the treatment group are approximately 22 hours, and middle-income in-work group are about 36 hours. The on-welfare group barely works.

Figure 1.3 provides children's scores by child age and single mother's working status, respectively. Panel (a) presents the cognitive test score gaps among three groups of lone mother status. There is a persistent cognitive test score gap between children of on-welfare mother and in-work mothers. For example, children of low-income in-work mothers(treatment group) have cognitive test scores around 0.16 standard deviation higher than those of on-welfare mothers, but 0.12 standard deviation lower than those of middle-income in-work mothers at age 3. The cognitive score gap between the low-income in-work and middle-income in-work mothers is relatively smaller than the gap between low-income in-work and low-income on-welfare group.

Panel (b) of figure 1.3 presents persistent non-cognitive score gaps among three groups of lone mother's working status. Children of low-income in-work mothers have 0.22 standard deviation higher than those of on-welfare families, but a 0.36 standard deviation lower than those of middle-income in-work mothers at child age 3. The non-cognitive skill gaps between the low-income in-work mothers and low-income on-welfare is also smaller then the gap between low-income in-work mothers and middle-income in-work mothers

Table 1.3 shows that children of on-welfare mothers always have the lowest cognitive test scores and the highest problematic behaviour compared to the in-work mothers. In general, low-income on-welfare mothers are less likely to use childcare arrangement, and middle-income in-work mothers slightly use more of formal childcare than low-income in-work mothers. Family learning environment measures the learning activity mother involved with child, low-income on-welfare mothers have the lowest score. Birthweight of cohort baby is monotonically increasing from low-income on-welfare group to middle-income in-work group. As for family structure, on-welfare group has significantly higher number of children in the household than in-work groups. When it comes to the education level, middle-income group has much higher proportion of mothers with first degree or above than the low-income groups.

Due to the changing definition of in-work benefit eligibility, table 1.4 shows the transition matrix of working status of mothers across the whole sample period. Panel a shows that most mothers stays in the same group across 4 waves. Among the mothers start as low-income non-working group, 81.68% of them stay in the same group and the majority of the rest (17.91%) move to low-income in-work group(eligible group). There are 75.32% of the low-income in-work group stays as eligible, 19.96% move to non-working group and 4.72% move to middle-income group. 73.30% of the middle-income in-work group stay in the same group, 23.30% move to the eligible group and 3.40% move to the on-welfare group. A simple regression (1.5) shows that mothers with more children are less likely to transition into eligible group or stay at the eligible group, while they are more likely to move to lower income and less working hour/out of labour market status. Mother's age at child birth is not correlated with transitioning into or transitioning out of middle-income in-work status, while older mothers tend to transition into or stay as the eligible group. Mothers with first degree are more likely to transition into middle-income group and are less likely to stay as low-income group. There is also evidence of regional effects found for London, South East and South West[1]. For instance, mothers living in London tend to move out of eligible group and transition into middle-income in-work group.

## 1.5 Empirical Results

This section firstly presents estimated effect of being entitled to in-work benefit on children's development for each specification and explores the reliability of each model for out-of-sample forecasting performance using cross-validation criterion. Then it provides various robustness check and tests possible channels through which the treatment group is affected.

### 1.5.1 Main Specifications

This subsection presents estimation results based on specifications mentioned in section 1.3: the contemporaneous specification, the cumulative specification, the value-added specification, value-added with IV specification and value-added augmented with cumulative specification. The intergenerational effect of being eligible to the in-work benefit is given by the coefficient on low-income in-work, with low-income on-welfare and middle-income in-work as reference groups, separately, conditional on demographic characteristics and regional fixed effects. Results on cognitive outcome and non-cognitive outcome refer to slightly different sample, due to the missing.

Table 1.6 reports the in-work reform effects on children's cognitive achievements. Column (1) shows the results of contemporaneous specification, which is the baseline model. It is found that mothers who receive in-work benefit are more likely to improve children's cognitive achievements than those mothers living on welfare, the treatment group increases cognitive outcome significantly by about 0.058 of a standard deviation(Panel a). Surprisingly, there is no significant effect found for the eligible group with middle-income in-work group as the reference group (Panel b). Nonetheless, the contemporaneous specification requires strong restrictions on the production function.

[1]There are in total 12 regions included in the regression. They are North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales, Scotland and Northern Ireland. The reference group is North east.

For this specification to make sense, it assumes that test scores are only determined by current working status conditional on observed characteristics.

To test the implication of contemporaneous specification, this paper includes historical status in column (2). Only lagged one period status is considered due to the data restriction. As indicated in column (2) of panel a, the estimated effects of current working status decreases to 0.044 of a standard deviation but the effects of past working status is not significant. Panel b shows an insignificant current treatment effect of 0.021 and past treatment effect of -0.018 of a standard deviation, comparing to the middle-income in-work group. The null hypothesis of cumulative specification is rejected, which implies the in favour of contemporaneous specification.

According to the value added specification in column (3), treatment improves cognitive outcome by 0.052 of a standard deviation comparing to the low-income non-work group, while no significant effect found comparing to the middle-income in-work group. In panel a, past test scores significantly increase children's cognitive outcome by 0.317 of a standard deviation. Panel b shows a significant effect of 0.262 of a standard deviation. It may imply that the past test scores capture the effects of omitted current and past inputs, and unobserved endowment or shocks. Note that this specification implicitly allows for measurement error in test scores and heterogeneity in children's learning ability. By instrumenting lagged one period outcome by lagged two period outcome, column (4) of panel a finds that the treatment effect rises from significant 0.052 in the value-added model to 0.079 and the persistence parameter increases from 0.317 to 0.739 of a standard deviation. This is consistent with equation 1.15 that attenuation bias exists in the value added specification. Similarly, panel b shows an increase in the persistence parameter, rising from 0.262 to 0.690 of a standard deviation. Notice there is an additional issue with the value-added model. Estimates of the persistence parameter are biased due to both measurement error in test scores and unobserved child-level heterogeneity in learning. As specified in equation 1.13, past test scores fully capture individual heterogeneity only if it enters through a one-time process, but this assumption will be violated if talented children also learn faster. The positive correlation between this unobserved heterogeneity and lagged past scores will lead to an upward bias of parameter $\theta$. Therefore, the direction of expected bias resulting from a simple value-added model depends on the relative magnitude of two sources of bias.

Column(5) shows the estimation result when controlling for both historical working status and and children's past cognitive outcomes. The effect of current working status decreases to insignificant 0.010 comparing to the low-income no-work group, while the lagged working status picks up the significant effect of 0.078. With the middle-income in-work group as the reference group, both current and past working status effect are insignificant. The persistence parameter on the lagged outcome is similar to the value added model.

In general these models show that girls have positive effects on cognitive test scores than boys, increasing outcome by 0.059 to 0.106 of a standard deviation in panel a, 0.079 to 0.112 of a standard deviation in panel b. Birthweight positively predicts test score, with association ranges from 0.029 to 0.120. In terms of family structure, more siblings not necessarily reduce the cognitive score, and higher birth order adversely affects outcome, ranging from 0.036 to 0.090. Higher-educated mothers improve children's cognitive outcome. Having a first degree or above considerably increases test score by 0.024 to 0.154 of a standard deviation in panel a and by 0.002 to 0.088 of a standard deviation in panel b.

Table 1.7 provides evidence of the in-work benefit effect on children's non-cognitive outcomes. Generally, the treatment effect on non-cognitive score is, in magnitude, greater than the effect on cognitive score. Under the contemporaneous and cumulative specification, mothers currently receiving tax credit have significantly positive effect on children's non-cognitive behaviour than mothers living on welfare(Panel a). For example, column (1) provides evidence that treatment group has non-cognitive outcome significantly higher by 0.152 of a standard deviation. The cumulative specification gives similar results on the effect of current working status, while the effect of past working status is insignificant and small in magnitude. The value-added model gives smaller treatment effect, 0.071 of a standard deviation. Instrumenting the lagged non-cognitive score increases the persistence parameter from 0.544 to 0.901 (column (4)). It seems that the current working status has no significant effect and it is absorbed by the effect of lagged non-cognitive score. The effect given by cumulative value-added specification is consistent with the value-added model, with no effect captured by lagged working status.

Departure from no effect on cognitive scores, treatment group shows significantly adverse effect on non-cognitive scores comparing to the middle-income in-work group, under some specifications(Panel b). The magnitude of treatment effect ranges from insignificant -0.002 to -0.224 of one standard deviation. While the past working status has no significant effect on non-cognitive outcomes, the persistence parameter is around 0.496, increasing to 1.074 of a standard deviation when the lagged test score is instrumented.

In general, girls have less problematic behaviours than boys, and birthweight positively predicts non-cognitive score, with association range from 0.018 to 0.080. In other words, children with one unit higher birthweight (which corresponding to one standard deviation higher) reduce problematic behaviour by 0.018 to 0.080 of a standard deviation. The number of siblings and birth parity seem to matter less for non-cognitive outcomes than cognitive outcomes. The effect of mother having a first degree or above considerably increase children's non-cognitive score. Ethnicity seems to have no effect on child's behavioural performance.

Overall, the estimation results differ by underlying model assumptions and outcome variables. However, the treatment effect shows an improvement on both child cognitive test scores and non-cognitive well-being, comparing to the low-income non-work group. With the middle-income in-work mothers as the reference group, the treatment effect is insignificant for cognitive outcome and negative for non-cognitive scores. Due to the possibility of selection into treatment and intrinsic differences between treatment and control groups, all the estimators give only suggestive evidences but not causal effects.

### 1.5.2 Model Selection

Given all these specifications and discussion of estimation results, an important issue is to select among competing specifications which have good representation of results. Referring to Todd and Wolpin (2003) and Del Bono et al. (2016), this paper evaluates and compares models by applying cross-validation method in addition to conventional specification tests. It is useful for testing among non-nested models and non strictly preferred null hypothesis. This method compares the performance of different models and select the one with the best goodness fit for the available data on the basis of an out-of-sample root-mean-squared error (RMSE) criterion.

This paper chooses the K-fold cross-validation which is the basic form of cross-validation to implement our selection. The process is as follows. First, the sample is partitioned into six nearly equally sized sub-samples. Note that there are three categories of single mothers on the sample, the data need to be stratified prior to being divided into six sub-samples in order to ensure that each sub-sample is a good representative of the whole. Subsequently six iterations of estimation and validation are performed for each model such that within each iteration five sub-samples are used for estimation while the remaining one sub-sample is held-out for validation. In other words, the estimation results from the five sub-samples are used to construct the RMSE for the one held-out sub-sample. In order to avoid coincidence, the RMSE is constructed from four different initial randomizations and then average the four RMSE for each specification.

Table 1.8 provides the RMSE for each specification and for both cognitive outcome and non-cognitive outcome models. For the first control group, the first column favours the cumulative value-added specification for the cognitive outcomes and the second column selects the value-added model for non-cognitive outcomes(panel a). The second control group chooses the value-added specification for both cognitive and non-cognitive outcomes(panel b).

Referring back to the estimation results shown in table 1.6 and table 1.7, the specification with the lowest RMSE suggests that lagged in-work benefit eligibility significantly increases children's cognitive outcomes by 0.078 of a standard deviation, and improves non-cognitive skills by 0.071 of a standard deviation, comparing to the mothers living on welfare.

### 1.5.3 Robustness Check

This subsection presents additional findings from following perspectives: alternative definition of groups; child fixed effects to allow employment to be endogenous with respect to unobserved endowment and time-invariant maternal characteristics; possible mechanisms; the consistency of two in-work benefit reforms on child outcomes. The robustness check is only focus on baseline specification and the specification with the lowest RMSE as selected in table 1.8.

#### 1.5.3.1 Alternative Definition of Groups

If mothers endogenously adjust labour supply with respect to changes in tax credit eligibility, then it would be preferred to define the treatment and control groups before the implementation of in-work benefit reform or define groups by variables that are not directly related to benefit eligibility. This subsection firstly defines groups based on the first wave of the available data, that is the pre 2003 WTC/CTC reform period. Thus the treatment group and control groups are selected based on the 2002-2003 welfare rule. For a lone parent family with characteristics at the sample means (1.78 children, £13.67 weekly childcare costs), tax credit reaches zero for a weekly net income of £314.79. In addition, education level would be a good candidate for defining treatment and control groups.

The size of the on-welfare group increases, the size of the treatment group increases in the second wave but decreases in the third and fourth wave, comparing to the sample size in the main regression. Figure 1.4 shows the labour supply status of treatment and control groups. The middle-income in-work group has lower income than that is shown in figure 1.2. Figure 1.5 presents child cognitive and non-cognitive outcomes by child age and groups, and there are persistent gaps across groups across time. Comparing with figure 1.3, the non-cognitive score gap between treatment group and middle-income in-work group is 0.18 of a standard deviation smaller at age 3, 0.05 smaller at age 5, and 0.13 smaller at age 7. The non-cognitive score gap is also smaller than shown in figure 1.3 between the treatment group and the non-welfare group, with 0.04 of a standard deviation smaller at age 5 and 0.07 at age 7.

Table 1.9 provides estimation results of the treatment effect. There is no significant effect found on the cognitive test scores under the baseline, irrespective of the control group. Value-added specification shows a significant 0.041 of a standard deviation increase in the cognitive test score, comparing to the low-income non-work group. This is similar to the estimation result of value-added specification in table 1.6. Non-cognitive score is significantly improved within the treatment group under the baseline specification, but the estimates significantly reduced under the value-added specification, with the second control group as the reference group. This is in contrast with the main regression result of a significant 0.071(table 1.6). One candidate explanation is the change in sample composition and smaller non-cognitive gap between the treatment and the second control group, when the treatment and control groups are fixed at the beginning period.

Table 1.10 offers evidence on an alternative definition of treatment and control groups—the education level. Specifically, the treatment group includes mothers who work at least 16 hours per week and have education level above O-level but below bachelor level. The reference group is comprised of mothers who work less than 16 hours per week with the education level above O-level but below bachelor level. The second control group is mothers who work no less than 16 hours per week and have at least bachelor education degree. Results are similar to table 1.9.

#### 1.5.3.2 Child Fixed Effects

Above main specifications include a set of demographic variables to control for potential differences in group-specific compositional changes over time. However, those models deal with the effects of unobserved time-invariant characteristics by making strong assumptions. Here I try to relax those assumptions by employing child fixed effects. It allows employment choice to be endogenous with respect to unobserved child endowment and compositional changes in unobserved characteristics.

Panel a of table 1.11 shows that baseline estimator gives a significant treatment effect of -0.105 on cognitive scores by using the first control group, the magnitude reduces to 0.041 once the preferred cumulated value added model controls for past working status and outcome. Surprisingly, the persistent parameter gives negative value for non-cognitive outcome in panel b. Under the child fixed effects specification, child endowment and unobserved time-invariant variables are no longer captured by past child outcome any more. The child fixed effects account for time-invariant differences between treatment and control groups, under the assumption that maternal employment is endogenous only with respect to time-invariant characteristics. Child fixed effect model fails to document any significant treatment effect in the preferred specification, with the on-welfare group as the reference group.

#### 1.5.3.3 Mechanism

In order to examine the mechanisms through which eligibility to in-work benefits might affect children's outcomes, this paper is particularly interested in the potential adjustments in usage of childcare, maternal time inputs, increased income and working hours (Duncan et al., 2000; Zaslow et al., 2002; Bernal and Keane, 2011). It has been well documented that in-work benefit reform encourages usage of child care service, and Bernal and Keane (2010) find that one more year of child care decreases child cognitive score by 2.1%. Both the WFTC and WTC/CTC policy contains an important childcare compensation parameter. Both table 1.12 and table 1.13 show that the treatment effect turn insignificant once controlling for childcare usage and maternal time inputs.

Column(1) and (3) of table 1.12 find that formal child care increases child cognitive outcome by 0.092 and improves children's emotional skill by 0.061 of a standard deviation, with respect to the first control group. Cumulative value-added models(column (2)) show that lagged formal child care significantly improves cognitive scores by 0.132 of a standard deviation, while no childcare effects found for non-cognitive outcomes under the value added model(column(4)). Table 1.13 shows no significant effect of formal childcare but a negative effect of informal childcare of -0.263 of a standard deviation for cognitive skill development under the baseline specification. There is no significant effect of childcare on children's non-cognitive skill. These evidence may suggest that treatment group improves children's outcome through more formal childcare usage, comparing to the low-income non-work group.

Del Bono et al. (2016) shows a positive relationship between maternal time inputs and child cognitive and non-cognitive outcomes. If being eligible to the in-work benefit adjusts mother's time allocation, then the treatment effect will partly work through the maternal time inputs. The MCS dataset provides a rich measurement of maternal time inputs(family learning environment) which foster children's development. At age three, mothers reported a range of activities that she spent with the child, including reading, going to the library, drawing, playing music, teaching numbers or counting, learning alphabet. These activities are reported in a frequency scale, ranging from"7 times a week" to"none at all". At age 5 and 7, the MCS provides more broad range of activities, including reading, going to the library, drawing, playing music, telling stories, doing sports, playing games/toys, going to the park. Based on the work of Del Bono et al. (2016), the indicator of maternal inputs is constructed by using principal component analysis. This study retains a single common factor which with eigenvalues greater than one according to the Kaiser criterion. This explains about 42.3 percent of the total variance at age 3, 32.6 percent at age 5 and 33.7 percent at age 7.

As it indicated in table 1.3, on-welfare group are less likely to interact with their children than the in-work mothers. Table 1.12 shows that family learning environment significantly improves children's cognitive score by 0.028 and non-cognitive skill by 0.096 of a standard deviation under the baseline specification. The effect reduces to insignificant 0.006 and significant 0.051, separately for cognitive and non-cognitive outcomes, under the RMSE-selected model. There is no effect of past family learning environment for children's outcome. Table 1.13 shows that lagged maternal time inputs have significant effect for children's non-cognitive outcomes.

Miller and Zhang (2009) and Dahl and Lochner (2012) exploit the effect of Earned Income Tax Credit(EITC) on the test scores of child and find that the improvement rate of children from low-income family are bigger than middle-income family. If this also applies to the context of UK welfare systems, then I would expect that children of lower end of income distribution of eligible group are more likely to be positively affected by the reforms, comparing to children of higher end of income distribution of eligible group. For the estimation of income effect, I explore the income and working hour effects by using interaction term rather than directly control for them since treatment status is defined based on income and working hours. The net annual income band that ranges from 2 (Less than £1,050) to 18 (£34,600 less than £52,000). Table 1.14 shows that the additional one more unit of income band is insignificantly associated with treatment group, irrespective of the reference group and specification. Thus it fails to provide any evidence on the income channel. Column (1) of table 1.15 shows that an additional hours of working reduces children's cognitive score by 0.013 of a standard deviation, comparing to the low-income non-work group. However, there is no evidence of working hour channel under the other specification and for non-cognitive outcomes.

To investigate further the effects of working hours on child outcome that are induced by eligibility to tax credit, the sample is split by working hours above or below 30 hours per week, which is another threshold for obtaining additional tax credit. Single claimants, which applies to the case of lone mothers, who work at least 30 hours per week are qualified to have the 30 hour element included in the award. Specifically, the low-income in-work group is divided into two subgroups—low-income in-work less than 30 hours group and low-income in-work no less than 30 hours group. Likewise, the middle-income in-work groups have been divided into two subgroups by working hours. According to the RMSE-selected models of table 1.16(panel a), lagged treatment group has significant effect on children's cognitive outcome, irrespective of working hours. In-work benefit eligible mothers who work less than 30 hours has significant effect on improving non-cognitive skills by 0.088 of a standard deviation, while treatment group with working hours more than 30 hours shows no significant effect. Table 1.16(panel b) show that treatment group has no significant effect on children's outcome under RMSE-selected models. Thus the positive treatment effects found in the main regression on non-cognitive score is partly explained by in-work benefit eligible mothers who work less than 30 hours per week.

#### 1.5.3.4 Difference-in-difference Estimations

In order to check the consistency of 1999 WFTC and 2003 WTC/CTC reform effect on children, a measure of pre 2003 WTC/CTC child outcome is needed. Table 1.6 and 1.7 show that child birth weight is a good predictor for later cognitive and non-cognitive outcomes. Thus I use standardized birthweight as a proxy for pre 2003 WTC/CTC child outcome. The sample is restricted to treatment group and the first control group—the low-income on-welfare mothers. A difference-in-difference approach is employed to check the effect of 2003 WTC/CTC reform on child outcome. Under the baseline framework, the difference-in-difference approach can be formalized as follows

$$Y_{it} = \phi_1 D_{it} + \phi_2 Post_{it} + \phi_3 D_{it} * Post_{it} + \phi_4 t + X_{it}\alpha_1 + e_{it} \tag{1.18}$$

where $Post_{it}$ is an indicator for post 2003 WTC/CTC reform, $\phi_3$ is the parameter of interest that captures the treatment effects. $\phi_2$ represents a common shift in the average value of $Y_{it}$ for both treatment and control groups. In other words, it captures the effects of all the other contemporaneous policies changes that occurred at the implementation of 2003 WTC/CTC. $\phi_4$ represents a linear time trend common to both groups. Column (1) and (4) of table 1.17 show that there is no significant difference in the treatment effect before and after the 2003 WTC/CTC reform.

Additionally, I relax the common trend assumption that is imposed on equation 1.18, to allow for child outcome of treatment group to evolve differently from those of control group regardless of the in-work benefit reform,

$$Y_{it} = \phi_1 D_{it} + \phi_2 Post_{it} + \phi_3 D_{it} * Post_{it} + (\phi_{41} + \phi_{42} D_{it})t + X_{it}\alpha_1 + e_{it} \tag{1.19}$$

where $\phi_{42}$ captures the differential time trend between treatment and control group. Column (2) and (5) of table 1.17 show no effect of policy changes on both child outcomes once conditional on different time trend.

Child fixed effects specification is also adopted to allow for the endogeneity between changes in mother's labour supply choices and the policy.

$$Y_{it} = \phi_1 D_{it} + \phi_2 Post_{it} + \phi_3 D_{it} * Post_{it} + (\phi_{41} + \phi_{42} D_{it})t + X_{it}\alpha_1 + C_i + e_{it} \tag{1.20}$$

where $C_i$ represents the child fixed effects. It takes advantage of the panel structure and allows for compositional changes in unobserved characteristics. Column (3) and (6) of table 1.17 display insignificant policy effects.

The results confirm the consistent effect of being eligible to 2003 WTC/CTC with 1999 WFTC on children's outcomes, therefore it should alleviate, to some extent, the concern of estimation bias that is resulting from treating eligibility to those two policies indifferently in this paper.

## 1.6 Conclusion

The Working Family Tax Credit(WFTC) replaced the Family Credit(FC) as the main source of in-work support for low-income families with dependent children in the UK in October 1999. In April 2003, Working Tax Credit and Child Tax Credit(WTC/CTC) then replaced the WFTC, while main features of WFTC have remained. This paper provides the first piece of evidence on the effect of being eligible to post 1999 in-work benefit policies on children's early stage development in Britain. To be specific, it studies the effect of the in-work benefit element of 1999 WFTC and 2003 WTC/CTC on children's outcomes.

It makes a number of improvements over previous literature. First, it finds that single mothers who are entitled to in-work benefit significantly improves children's cognitive outcome and non-cognitive outcome at early childhood stage. Second, it identifies treatment and control groups by exploring eligibility of benefit rules. In particular, since there is no perfect control group, this paper employs two control groups—one is more similar to the treatment group on one dimension that they have low income and the other is more similar in another dimension that they participate in the labour market. The advantage of having multiple control groups is that comparisons between different group pairs are likely to provide a bound estimator in the sense that it bounds the income effects and working hour effects. Additionally I can be more confident in that estimators are less biased by the effect of other contemporaneous changes or trend differences among groups. Third, various specifications of production function are used. Treatment effects are implicitly proxy for family monetary and time investment. Then a cross-validation method selects the value-added specification with the best performance as the main specification in this paper. Finally, this study explores the panel nature of the data and identifies which parameters of the in-work benefit policies are more likely to explain its estimated effects.

This paper stresses six main findings. First, the preferred value-added specification indicates that mothers who are eligible to in-work benefits positively affect children's cognitive scores by 0.078 of a standard deviation and non-cognitive outcomes by 0.071 of a standard deviation. Second, this paper finds no income effect that is induced by in-work benefit eligibility. It is in contrast with the finding of Miller and Zhang (2009) and Dahl and Lochner (2012) that document significant policy induced income effects. Third, increasing formal childcare usage induced by in-work benefit policy helps explain the positive treatment effect on cognitive outcomes, and increasing maternal time inputs contribute to an improvement in children's non-cognitive outcomes. Fourth, the improvement in non-cognitive skill is mainly driven by children of mothers who work less than 30 hours per week. The positive treatment effect fades when mother works at least 30 hours. It provides some implication for policy makers on designing the policy parameter. Fifth, by allowing mother's employment choice to be endogenous with respect to unobserved child endowment and compositional changes in unobserved characteristics, a child fixed effect model suggests no evidence of treatment effects. Sixth, this paper does not distinguish between the intergenerational effects of 1999 WFTC and 2003 WTC/CTC, for the reason that the changes in policy parameter are unlikely to affect working status of lone mothers and this study confirms the consistency in the effects of being entitle to 2003 WTC/CTC with 1999 WFTC on children's outcomes under a difference-in-difference framework.

In-work benefit policies aim to encouraging labour market participation of lone mothers and reducing child poverty. However, the intergenerational consequences of those policies on child outcomes are often overlooked. This paper claims that there is positive effect of in-work benefit eligibility on both children's cognitive and non-cognitive outcomes, and the positive impacts on non-cognitive outcome is mainly driven by treatment mother who works less than 30 hours per week. However, a comprehensive evaluation must also take into account other contemporary policy changes and other responses. The author is also aware that the underlying single mother model may not be sufficient in the case that they also receive other income support from boyfriend, ex-husband and extended family members.

# Tables

Table 1.1: Working Family Tax Credit (WFTC) Rule

| | 1999-2000 | 2000-01 | 2001-02 | 2002-03 |
|---|---|---|---|---|
| **Basic Credit** | | | | |
| Adult (couple/lone parent) | 52.30 | 53.15 | 59.00 | 62.50 |
| 30 hour premium | 11.05 | 11.25 | 11.45 | 11.65 |
| **Child Element** | | | | |
| Aged 0-11 | 19.85 | 25.60 | 26.00 | 26.45 |
| Aged 11-16 | 20.90 | 25.60 | 26.00 | 26.45 |
| Aged 16-18 | 25.95 | 26.35 | 26.75 | 27.20 |
| Maximum eligible childcare expenditure | 70.00 | 70.00 | 135.00 | 135.00 |
| Maximum eligible childcare expenditure, 2 or more children | 105.00 | 105.00 | 200.00 | 200.00 |
| Childcare subsidy | 70% | 70% | 70% | 70% |
| **Applicable Amounts** | | | | |
| Couple/Lone Parent | 90.00 | 91.45 | 92.90 | 94.50 |
| Wihtdrawal rate | 55% | 55% | 55% | 55% |

Source: Child Poverty Action Group, Welfare Benefits Handbook (various years).
Note: The tax credits are at weekly basis. Working Families' Tax Credit(WFTC) was introduced in October 1999 to replace Family Credit(FC). It was then integrated into the Child Tax Credit and Working Tax Credit(WTC/CTC) in April 2003.

Table 1.2: Working Tax Credit and Child Tax Credit (WTC/CTC) Rule

| | 2003-04 | 2004-05 | 2005-06 | 2006-07 | 2007-08 |
|---|---|---|---|---|---|
| **Working Tax Credit** | | | | | |
| Basic Element | 29.33 | 30.19 | 31.15 | 32.02 | 33.27 |
| Extra for couples and lone parents | 28.85 | 29.71 | 30.67 | 31.54 | 32.69 |
| 30 hour element | 11.92 | 12.31 | 12.69 | 13.08 | 13.56 |
| Disabled worker element | 39.23 | 40.38 | 41.63 | 42.79 | 44.42 |
| Severe Disability element | 16.63 | 17.12 | 17.69 | 18.17 | 18.85 |
| Maximum eligible childcare expenditure, 1 child | 135 | 135 | 175 | 175 | 175 |
| Maximum eligible childcare expenditure, 2 or more children | 200 | 200 | 300 | 300 | 300 |
| Proportion of eligible childcare covered | 70% | 70% | 70% | 80% | 80% |
| **Child Tax Credit** | | | | | |
| Family element | 10.48 | 10.48 | 10.48 | 10.48 | 10.48 |
| Family element, extra for child under 12 months | 10.48 | 10.48 | 10.48 | 10.48 | 10.48 |
| Child element | 27.79 | 31.25 | 32.50 | 33.94 | 35.48 |
| Disabled child additional element | 41.44 | 42.60 | 43.94 | 45.19 | 46.92 |
| Severely child additional element | 16.63 | 17.12 | 17.69 | 18.17 | 18.85 |
| **Common Elements** | | | | | |
| Threshold | 97.30 | 97.30 | 97.30 | 100.38 | 100.38 |
| Threshold if not entitled to WTC | 254.42 | 259.23 | 267.50 | 272.21 | 278.75 |
| Withdrawal rate | 37% | 37% | 37% | 37% | 37% |

Source: HM Revenue and Customs, http://www.hmrc.gov.uk/rates/taxcredits.htm
Note: For ease of comparison with WFTC rule, the tax credits in this table are imputed at weekly basis using annualised amounts.

Table 1.3: Summary Statistics by Child Age and Group

| | age 3 | | | age 5 | | | age 7 | | |
|---|---|---|---|---|---|---|---|---|---|
| | low welfare | low work | middle work | low welfare | low work | middle work | low welfare | low work | middle work |
| **Outcome variable** | | | | | | | | | |
| cognitive score | -0.031 | 0.130 | 0.246 | -0.124 | 0.015 | 0.091 | -0.289 | -0.0838 | 0.075 |
| | (0.555) | (0.526) | (0.628) | (0.585) | (0.523) | (0.589) | (0.747) | (0.612) | (0.572) |
| non-cognitive score | -0.318 | -0.100 | 0.260 | -0.321 | -0.112 | 0.118 | -0.382 | -0.133 | 0.116 |
| | (0.790) | (0.638) | (0.528) | (0.834) | (0.633) | (0.588) | (0.889) | (0.698) | (0.543) |
| **Childcare arrangement** | | | | | | | | | |
| none | 0.176 | 0.088 | 0.000 | 0.705 | 0.016 | 0.013 | 0.693 | 0.009 | 0.000 |
| formal | 0.715 | 0.304 | 0.365 | 0.055 | 0.220 | 0.377 | 0.057 | 0.207 | 0.368 |
| informal | 0.096 | 0.159 | 0.027 | 0.217 | 0.396 | 0.312 | 0.224 | 0.439 | 0.347 |
| family learning environment | -0.047 | 0.131 | 0.095 | -0.011 | 0.058 | 0.095 | -0.039 | 0.050 | 0.069 |
| | (1.020) | (0.939) | (1.026) | (1.042) | (0.964) | (0.766) | (1.061) | (0.916) | (0.859) |
| **Demographic charaterastics** | | | | | | | | | |
| child gender | 0.488 | 0.518 | 0.486 | 0.490 | 0.512 | 0.455 | 0.491 | 0.502 | 0.516 |
| birthweight | 3.076 | 3.140 | 3.244 | 3.068 | 3.143 | 3.219 | 3.064 | 3.129 | 3.253 |
| | (0.575) | (0.560) | (0.520) | (0.573) | (0.571) | (0.466) | (0.567) | (0.581) | (0.465) |
| no. of siblings | 1.176 | 0.670 | 0.703 | 1.401 | 0.865 | 0.792 | 1.668 | 0.986 | 0.884 |
| | (1.124) | (0.895) | (0.903) | (1.180) | (0.881) | (0.937) | (1.214) | (0.916) | (0.886) |
| parity | 1.894 | 1.544 | 1.527 | 1.895 | 1.611 | 1.571 | 1.924 | 1.616 | 1.600 |
| mother age | 25.470 | 26.860 | 32.340 | 25.340 | 26.800 | 32.260 | 25.280 | 26.530 | 32.150 |
| | (6.374) | (5.938) | (5.562) | (6.349) | (6.072) | (5.384) | (6.293) | (6.138) | (5.702) |
| dummy for having first degree or above | 0.023 | 0.073 | 0.541 | 0.018 | 0.073 | 0.519 | 0.018 | 0.058 | 0.474 |
| white | 0.895 | 0.930 | 0.878 | 0.892 | 0.932 | 0.844 | 0.887 | 0.935 | 0.832 |
| | (0.307) | (0.255) | (0.329) | (0.310) | (0.253) | (0.365) | (0.316) | (0.246) | (0.376) |
| **Region** | | | | | | | | | |
| North West | 0.107 | 0.088 | 0.162 | 0.106 | 0.095 | 0.143 | 0.106 | 0.101 | 0.105 |
| Yorkshire and the Humber | 0.062 | 0.079 | 0.000 | 0.065 | 0.067 | 0.026 | 0.065 | 0.066 | 0.053 |
| East Midlands | 0.046 | 0.055 | 0.041 | 0.052 | 0.038 | 0.065 | 0.043 | 0.050 | 0.063 |
| West Midlands | 0.071 | 0.068 | 0.041 | 0.070 | 0.070 | 0.052 | 0.073 | 0.066 | 0.053 |
| East of England | 0.054 | 0.050 | 0.041 | 0.053 | 0.060 | 0.039 | 0.055 | 0.056 | 0.063 |
| London | 0.091 | 0.048 | 0.189 | 0.088 | 0.055 | 0.169 | 0.088 | 0.055 | 0.168 |
| South East | 0.084 | 0.099 | 0.081 | 0.083 | 0.096 | 0.091 | 0.081 | 0.091 | 0.168 |
| South West | 0.044 | 0.051 | 0.041 | 0.041 | 0.057 | 0.052 | 0.036 | 0.061 | 0.032 |
| Wales | 0.210 | 0.181 | 0.149 | 0.217 | 0.179 | 0.130 | 0.218 | 0.190 | 0.095 |
| Scotland | 0.106 | 0.148 | 0.135 | 0.101 | 0.151 | 0.130 | 0.103 | 0.142 | 0.116 |
| Northern Ireland | 0.082 | 0.099 | 0.081 | 0.084 | 0.092 | 0.078 | 0.086 | 0.088 | 0.074 |
| **Observations** | 1386 | 546 | 74 | 1242 | 687 | 77 | 1106 | 805 | 95 |

Note: This table provides the descriptive statistics for three groups across child development age 3, 5 and 7.
low-welfare is the first control group, which includes children of single mother who is not working or working less than 16 hours per week; low-work is the treatment group, which is consists of children of single mother who is working no less than 16 hours per week but with low income; middle-work is the second control group, which is comprised of children of mother who is working more or equal to 16 hours per week with middle income.
Standard deviation from mean is reported in parenthesis.

Table 1.4: Transition Matrix of Working Status across Time

| | Working status | | | |
|---|---|---|---|---|
| | Non-working control group 1 | Low-income in-work treatment group | Middle-income in-work control group 2 | No. |
| | | Panel a: 9 months to age 7 | | |
| Non-working | 81.68% | 17.91% | 0.41% | 4159 |
| Low-income in-work | 19.96% | 75.32% | 4.72% | 1653 |
| Middle-income in-work | 3.40% | 23.30% | 73.30% | 206 |
| No. | 3734 | 2038 | 246 | |
| | | Panel b: 9 months to age 3 | | |
| Non-working | 82.56% | 16.85% | 0.59% | 1531 |
| Low-income in-work | 28.57% | 64.29% | 7.14% | 420 |
| Middle-income in-work | 3.64% | 32.73% | 63.64% | 55 |
| No. | 1386 | 546 | 74 | |
| | | Panel c: age 3 to age 5 | | |
| Non-working | 81.96% | 17,825 | 0.22% | 1386 |
| Low-income in-work | 18.86% | 77.29% | 3.85% | 546 |
| Middle-income in-work | 4.05% | 24.32% | 71.62% | 74 |
| No. | 1242 | 687 | 77 | |
| | | Panel d: age 5 to age 7 | | |
| Non-working | 80.27% | 19.32% | 0.40% | 1242 |
| Low-income in-work | 15.57% | 80.49% | 3.93% | 687 |
| Middle-income in-work | 2.60% | 15.58% | 81.82% | 77 |
| No. | 1106 | 805 | 95 | |

Note: Panel a shows the transition matrix across the whole sample period. Panel b, panel c and panel d of the transition matrix show the probability of transiting between working status from wave $t(t <= 3)$ to wave $t+1$. There are three working status at each wave, with the low-income in-work as the treatment group status, low-income non-working and middle-income in-work as two control group status.

Table 1.5: The Determinates of Working Status' Transition

| | (1) status11[1] | (2) status22 | (3) status33 | (4) status12 | (5) status13 | (6) status23 | (7) status21 | (8) status31 | (9) status32 |
|---|---|---|---|---|---|---|---|---|---|
| child gender | -0.015 | -0.009 | -0.063 | 0.015 | 0.000 | -0.002 | 0.012 | -0.010 | 0.073 |
| | (0.012) | (0.021) | (0.066) | (0.012) | (0.002) | (0.010) | (0.019) | (0.030) | (0.065) |
| birth weight | -0.018 | 0.028 | -0.056 | 0.015 | 0.003 | 0.004 | -0.033 | 0.002 | 0.053 |
| | (0.010) | (0.019) | (0.065) | (0.010) | (0.002) | (0.009) | (0.017) | (0.030) | (0.064) |
| no. of children | 0.066*** | -0.101*** | -0.207*** | -0.064*** | -0.001 | -0.010 | 0.111*** | 0.063* | 0.144** |
| | (0.008) | (0.018) | (0.055) | (0.008) | (0.001) | (0.009) | (0.017) | (0.025) | (0.053) |
| birth order | -0.021* | 0.051* | 0.318*** | 0.021* | 0.000 | -0.008 | -0.042* | -0.061* | -0.257*** |
| | (0.010) | (0.021) | (0.063) | (0.010) | (0.002) | (0.010) | (0.019) | (0.029) | (0.062) |
| mother age at child birth | -0.023** | 0.057*** | 0.079 | 0.024** | -0.001 | -0.004 | -0.053*** | -0.015 | -0.064 |
| | (0.008) | (0.014) | (0.051) | (0.008) | (0.001) | (0.007) | (0.013) | (0.024) | (0.050) |
| mother age square | 0.000** | -0.001*** | -0.001 | -0.000** | 0.000 | 0.000 | 0.001*** | 0.000 | 0.001 |
| | (0.000) | (0.000) | (0.001) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.001) |
| dummy for having first degree or above | -0.141*** | -0.035 | 0.282*** | 0.110*** | 0.031*** | 0.107*** | -0.072** | -0.057 | -0.225** |
| | (0.027) | (0.029) | (0.073) | (0.027) | (0.005) | (0.014) | (0.027) | (0.033) | (0.071) |
| white | -0.036 | 0.089* | 0.022 | 0.034 | 0.002 | -0.040 | -0.049 | -0.034 | 0.011 |
| | (0.022) | (0.044) | (0.094) | (0.022) | (0.004) | (0.022) | (0.040) | (0.043) | (0.091) |
| Region Effects[a] | | | | | | | | | |
| London | -0.005 | -0.156* | -0.057 | -0.005 | 0.009 | 0.077* | 0.079 | 0.060 | -0.003 |
| | (0.037) | (0.072) | (0.184) | (0.037) | (0.006) | (0.035) | (0.066) | (0.084) | (0.180) |
| South East | -0.033 | -0.063 | -0.391* | 0.031 | 0.003 | 0.060 | 0.002 | 0.020 | 0.371 |
| | (0.035) | (0.063) | (0.194) | (0.035) | (0.006) | (0.031) | (0.058) | (0.089) | (0.189) |
| South West | -0.034 | 0.082 | -0.416* | 0.033 | 0.001 | -0.001 | -0.080 | 0.110 | 0.306 |
| | (0.041) | (0.070) | (0.203) | (0.040) | (0.007) | (0.035) | (0.064) | (0.093) | (0.198) |
| **Observations** | 4137 | 1651 | 206 | 4137 | 4137 | 1651 | 1651 | 206 | 206 |
| **R-squared** | 0.039 | 0.070 | 0.272 | 0.035 | 0.019 | 0.078 | 0.095 | 0.092 | 0.239 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level.

[1] There are three working status of mothers, non-working, low-income in-work and middle-income in-work. For the ease of presentation, number 1 indicates non-working group, number 2 represents low-income in-work group and number 3 means middle-income in-work group. Thus variable name "status11" indicates for mother's working status stays as non-working, "status12" means mother's working status transits from non-working to low-income in-work, the same logic applies to other dependent variables.

[a] There are 12 regions controlled for, with the North East region as the reference group. This table only presents regions with some significance of estimates.

Table 1.6: The Effect of In-work Benefit Reforms on Children's Cognitive Outcomes

| | (1) baseline | (2) cumulative | (3) VA | (4) VA-IV | (5) CVA[1] |
|---|---|---|---|---|---|
| | **Panel a: Control Group: low-income non-work** | | | | |
| **Working status**[a] | | | | | |
| low-income in-work | 0.058** | 0.044* | 0.052* | 0.079* | 0.010 |
| | (0.019) | (0.021) | (0.020) | (0.035) | (0.024) |
| lagged low-income in-work | | 0.029 | | | 0.078** |
| | | (0.021) | | | (0.026) |
| **Lagged child outcome** | | | 0.317*** | 0.739*** | 0.316*** |
| | | | (0.021) | (0.106) | (0.021) |
| **Demographic controls** | | | | | |
| child gender | 0.106*** | 0.105*** | 0.060** | 0.104** | 0.059** |
| | (0.019) | (0.019) | (0.019) | (0.033) | (0.019) |
| birth weight | 0.071*** | 0.070*** | 0.056*** | 0.120*** | 0.054** |
| | (0.018) | (0.018) | (0.017) | (0.029) | (0.017) |
| no. of siblings | -0.057*** | -0.057*** | -0.010 | 0.004 | -0.008 |
| | (0.015) | (0.015) | (0.014) | (0.021) | (0.014) |
| birth order | -0.056** | -0.055** | -0.068*** | -0.036 | -0.066*** |
| | (0.018) | (0.018) | (0.016) | (0.028) | (0.016) |
| mother age at child birth | 0.024 | 0.023 | 0.032* | -0.009 | 0.029* |
| | (0.013) | (0.013) | (0.013) | (0.023) | (0.013) |
| mother age square | -0.000 | -0.000 | -0.000* | 0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| dummy for having first degree or above | 0.154*** | 0.151*** | 0.126*** | 0.024 | 0.119*** |
| | (0.038) | (0.038) | (0.033) | (0.065) | (0.033) |
| white | 0.010 | 0.010 | -0.064 | -0.140* | -0.064 |
| | (0.041) | (0.041) | (0.037) | (0.063) | (0.037) |
| **Observations** | 5533 | 5533 | 3613 | 1731 | 3613 |
| **R-squared** | 0.067 | 0.067 | 0.146 | 0.035 | 0.148 |
| | **Panel b: Control Group: middle-income in-work** | | | | |
| **Working status**[b] | | | | | |
| low-income in-work | 0.012 | 0.021 | 0.049 | 0.034 | 0.064 |
| | (0.050) | (0.051) | (0.046) | (0.077) | (0.065) |
| lagged low-income in-work | | -0.018 | | | -0.024 |
| | | (0.050) | | | (0.066) |
| **Lagged child outcome** | | | 0.262*** | 0.690** | 0.262*** |
| | | | (0.038) | (0.226) | (0.038) |
| **Demographic controls** | | | | | |
| child gender | 0.112*** | 0.112*** | 0.079* | 0.097 | 0.080* |
| | (0.031) | (0.031) | (0.031) | (0.050) | (0.031) |
| birth weight | 0.059* | 0.059* | 0.029 | 0.074 | 0.029 |
| | (0.029) | (0.029) | (0.028) | (0.046) | (0.028) |
| no. of siblings | -0.016 | -0.016 | 0.013 | 0.045 | 0.013 |
| | (0.028) | (0.028) | (0.027) | (0.037) | (0.027) |
| birth order | -0.079* | -0.078* | -0.090** | -0.044 | -0.090** |
| | (0.033) | (0.033) | (0.032) | (0.049) | (0.032) |
| mother age at child birth | 0.003 | 0.003 | -0.007 | -0.016 | -0.007 |
| | (0.021) | (0.021) | (0.021) | (0.035) | (0.021) |
| mother age square | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) | (0.001) | (0.000) |
| dummy for having first degree or above | 0.088* | 0.087* | 0.087* | 0.002 | 0.085* |
| | (0.041) | (0.041) | (0.039) | (0.073) | (0.039) |
| white | 0.001 | 0.001 | -0.054 | -0.034 | -0.054 |
| | (0.067) | (0.067) | (0.067) | (0.092) | (0.067) |
| **Observations** | 1489 | 1489 | 1127 | 610 | 1127 |
| **R-squared** | 0.059 | 0.059 | 0.125 | . | 0.125 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level. All regressions control for region fixed effects.

[1] It represents the cumulative value added specification.

[a] The reference group is low-income on-welfare families, and low-income in-work is the treatment group.

[b] The reference group is middle-income in-work families, and low-income in-work is the treatment group.

Table 1.7: The Effect of In-work Benefit Reforms on Children's Non-cognitive Outcomes

| | (1) baseline | (2) cumulative | (3) VA | (4) VA-IV | (5) CVA[1] |
|---|---|---|---|---|---|
| | **Panel a: Control Group: low-income non-work** | | | | |
| **Working status**[a] | | | | | |
| low-income in-work | 0.152*** | 0.150*** | 0.071** | 0.044 | 0.092** |
| | (0.025) | (0.026) | (0.022) | (0.035) | (0.028) |
| lagged low-income in-work | | 0.004 | | | -0.039 |
| | | (0.026) | | | (0.029) |
| **Lagged child outcome** | | | 0.544*** | 0.901*** | 0.544*** |
| | | | (0.020) | (0.057) | (0.020) |
| **Demographic controls** | | | | | |
| child gender | 0.182*** | 0.182*** | 0.114*** | 0.089* | 0.115*** |
| | (0.029) | (0.029) | (0.021) | (0.035) | (0.021) |
| birth weight | 0.048* | 0.048* | 0.050** | 0.018 | 0.051** |
| | (0.024) | (0.024) | (0.018) | (0.028) | (0.018) |
| no. of siblings | -0.006 | -0.006 | -0.021 | 0.021 | -0.022 |
| | (0.021) | (0.021) | (0.016) | (0.023) | (0.016) |
| birth order | -0.035 | -0.035 | 0.015 | 0.016 | 0.014 |
| | (0.026) | (0.026) | (0.019) | (0.027) | (0.019) |
| mother age at child birth | 0.055** | 0.055** | 0.015 | -0.001 | 0.017 |
| | (0.019) | (0.019) | (0.016) | (0.021) | (0.016) |
| mother age square | -0.001* | -0.001* | -0.000 | 0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| dummy for having first degree or above | 0.243*** | 0.242*** | 0.114*** | -0.002 | 0.117*** |
| | (0.041) | (0.041) | (0.031) | (0.053) | (0.031) |
| white | -0.059 | -0.059 | -0.062 | -0.060 | -0.062 |
| | (0.049) | (0.049) | (0.041) | (0.067) | (0.042) |
| **Observations** | 5580 | 5580 | 3655 | 1775 | 3655 |
| **R-squared** | 0.054 | 0.054 | 0.309 | 0.262 | 0.309 |
| | **Panel b: Control Group: middle-income in-work** | | | | |
| **Working status**[b] | | | | | |
| low-income in-work | -0.183*** | -0.224*** | -0.078 | -0.002 | -0.137* |
| | (0.052) | (0.055) | (0.044) | (0.062) | (0.055) |
| lagged low-income in-work | | 0.075 | | | 0.096 |
| | | (0.056) | | | (0.058) |
| **Lagged child outcome** | | | 0.496*** | 1.074*** | 0.498*** |
| | | | (0.039) | (0.113) | (0.039) |
| **Demographic controls** | | | | | |
| child gender | 0.204*** | 0.204*** | 0.110*** | -0.054 | 0.109*** |
| | (0.040) | (0.040) | (0.032) | (0.051) | (0.032) |
| birth weight | 0.080* | 0.080* | 0.078** | -0.028 | 0.078** |
| | (0.036) | (0.036) | (0.027) | (0.045) | (0.027) |
| no. of siblings | 0.025 | 0.028 | -0.002 | -0.008 | 0.000 |
| | (0.039) | (0.039) | (0.031) | (0.044) | (0.031) |
| birth order | -0.036 | -0.038 | -0.010 | 0.005 | -0.012 |
| | (0.046) | (0.046) | (0.036) | (0.048) | (0.036) |
| mother age at child birth | 0.038 | 0.038 | 0.033 | 0.047 | 0.033 |
| | (0.028) | (0.028) | (0.025) | (0.030) | (0.025) |
| mother age square | -0.000 | -0.000 | -0.001 | -0.001 | -0.001 |
| | (0.000) | (0.000) | (0.000) | (0.001) | (0.000) |
| dummy for having first degree or above | 0.096 | 0.102 | 0.071 | -0.023 | 0.079* |
| | (0.052) | (0.052) | (0.040) | (0.064) | (0.040) |
| white | -0.050 | -0.051 | -0.079 | -0.180 | -0.079 |
| | (0.082) | (0.082) | (0.070) | (0.100) | (0.070) |
| **Observations** | 1510 | 1510 | 1150 | 633 | 1150 |
| **R-squared** | 0.088 | 0.088 | 0.296 | 0.191 | 0.298 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level. All regressions control for region fixed effects.

[1] It represents the cumulative value added specification.

[a] The reference group is low-income on-welfare families, and low-income in-work is the treatment group.

[b] The reference group is middle-income in-work families, and low-income in-work is the treatment group.

Table 1.8: Model Selection—A Cross-Validation Approach

| | (1)<br>Cognitive outcome | (2)<br>Non-cognitive outcome |
|---|---|---|
| | **Panel a: Control Group: low-income non-work** | |
| Contemporary | 3.5790 | 4.5848 |
| Cumulative | 3.5790 | 4.5826 |
| VA | 3.5125 | 3.9849* |
| VA-IV | 3.5393 | 4.0065 |
| CVA | 3.5084* | 3.9854 |
| | **Panel b: Control Group: middle-income in-work** | |
| Contemporary | 3.3169 | 3.6891 |
| Cumulative | 3.3169 | 3.6877 |
| VA | 3.2569* | 3.2740* |
| VA-IV | 3.2741 | 3.3155 |
| CVA | 3.2590 | 3.2769 |

Notes: This paper compares different specifications by applying cross-validation method. It compares the performance of different models and select the one with the best goodness fit for the available data on the basis of an out-of-sample root-mean-squared error (RMSE) criterion. This table applies the K-fold cross-validation.
* indicates the model specification with the smallest RMSE value.

Table 1.9: Robustness Check: Alternative Definition of Groups—Fixing groups at the first wave[1]

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Cognitive outcome | | Non-cognitive outcome | |
| | baseline | value added[2] | baseline | value added |
| **Panel a: Control Group: low-income non-work** | | | | |
| low-income in-work [a] | 0.023 | 0.041* | 0.105*** | 0.036 |
| | (0.020) | (0.019) | (0.027) | (0.020) |
| lagged cognitive score | | 0.314*** | | |
| | | (0.019) | | |
| lagged non-cognitive score | | | | 0.533*** |
| | | | | (0.018) |
| R-squared | 0.065 | 0.142 | 0.051 | 0.298 |
| Observations | 6786 | 4439 | 6837 | 4487 |
| **Panel b: Control Group: middle-income in-work** | | | | |
| low-income in-work[b] | -0.082 | -0.017 | -0.050 | -0.002 |
| | (0.050) | (0.042) | (0.056) | (0.045) |
| lagged cognitive score | | 0.274*** | | |
| | | (0.031) | | |
| lagged non-cognitive score | | | | 0.491*** |
| | | | | (0.037) |
| R-squared | 0.072 | 0.140 | 0.076 | 0.294 |
| Observations | 2003 | 1312 | 2024 | 1336 |
| Demographic controls [c] | YES | YES | YES | YES |
| Regional FE | YES | YES | YES | YES |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level.

[1] Define treatment and control groups based on the eligibility at the first wave of observation and keep the group composition the same across waves.

[2] The cross validation selects cumulative value-added specification for modelling the cognitive outcome for the first control group, but there is no variation in the working status in the case of fixing treatment group. Therefore, value-added model, the specification with the second smallest RMSE, is used here for the first control group.

[a] The reference group is low-income on-welfare families, and low-income in-work is the treatment group.

[b] The reference group is middle-income in-work families, and low-income in-work is the treatment group.

[c] Demographic controls include child gender, birthweight, number of siblings, birth parity, mother's age and age square, dummy for having first degree or above and ethnicity.

Table 1.10: Robustness Check: Alternative Definition of Groups—Education[1]

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Cognitive outcome | | Non-cognitive outcome | |
| | baseline | value added | baseline | value added |
| **Panel a: Control Group: low-income non-work** | | | | |
| low-income in-work[a] | 0.031 | 0.041* | 0.110*** | 0.038 |
| | (0.020) | (0.019) | (0.027) | (0.020) |
| lagged cognitive score | | 0.320*** | | |
| | | (0.019) | | |
| lagged non-cognitive score | | | | 0.537*** |
| | | | | (0.018) |
| R-squared | 0.057 | 0.136 | 0.042 | 0.296 |
| Observations | 6733 | 4405 | 6783 | 4451 |
| **Panel b: Control Group: middle-income in-work** | | | | |
| low-income in-work[b] | -0.092 | -0.110* | -0.119 | -0.055 |
| | (0.065) | (0.052) | (0.084) | (0.059) |
| lagged cognitive score | | 0.292*** | | |
| | | (0.029) | | |
| lagged non-cognitive score | | | | 0.494*** |
| | | | | (0.036) |
| R-squared | 0.070 | 0.149 | 0.069 | 0.298 |
| Observations | 2181 | 1431 | 2205 | 1458 |
| Demographic controls [c] | YES | YES | YES | YES |
| Regional FE | YES | YES | YES | YES |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level.

[1] Define treatment and control groups based on the education level of lone mother at the first wave of observation and keep the group composition the same across waves.

[a] The reference group is comprised of mothers who work less than 16 hours per week with the education level above O-level but below bachelor level. The treatment group includes mothers who work at least 16 hours per week and have education level above O-level but below bachelor level.

[b] The reference group is mothers who work no less than 16 hours per week and have at least bachelor education degree.

[c] Demographic controls include child gender, birthweight, number of siblings, birth parity, mother's age and age square, dummy for having first degree or above and ethnicity.

Table 1.11: Robustness Check: Child Fixed Effects Estimation

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Cognitive outcome | | Non-cognitive outcome | |
| | baseline | CVA[1] | baseline | VA |
| **Panel a: Control Group: low-income non-work** | | | | |
| low-income in-work[a] | -0.105*** | 0.041 | 0.029 | 0.023 |
| | (0.025) | (0.037) | (0.026) | (0.034) |
| lagged low-income in-work | | 0.018 | | |
| | | (0.034) | | |
| lagged cognitive score | | 0.292*** | | |
| | | (0.024) | | |
| lagged non-cognitive score | | | | 0.539*** |
| | | | | (0.021) |
| R-squared | 0.016 | 0.148 | 0.005 | 0.310 |
| Observations | 5533 | 2625 | 5580 | 3094 |
| | Cognitive outcome | | Non-cognitive outcome | |
| | baseline | VA | baseline | VA |
| **Panel b: Control Group: middle-income in-work** | | | | |
| low-income in-work[b] | 0.012 | 0.059 | -0.154* | -0.049 |
| | (0.062) | (0.057) | (0.071) | (0.072) |
| lagged cognitive score | | 0.244*** | | |
| | | (0.046) | | |
| lagged non-cognitive score | | | | -0.377*** |
| | | | | (0.046) |
| R-squared | 0.001 | 0.122 | 0.009 | 0.166 |
| Observations | 1489 | 768 | 1510 | 1150 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level. Additional controls includes number of siblings and region dummies.
[1] It represents the RMSE-selected cumulative value added specification.
[a] The reference group is low-income on-welfare families, and low-income in-work is the treatment group.
[b] The reference group is middle-income in-work families, and low-income in-work is the treatment group.

Table 1.12: Mechanism: Treatment Effect through Childcare Usage and Family Environment–First Control Group

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | cognitive | | non-cognitive | |
| | baseline | CVA | baseline | VA |
| **Working Status**[a] | | | | |
| low-income in-work | 0.045 | 0.041 | 0.100** | 0.023 |
| | (0.024) | (0.037) | (0.031) | (0.034) |
| lagged low-income in-work | | 0.018 | | |
| | | (0.034) | | |
| **Childcare arrangement**[b] | | | | |
| formal childcare | 0.092*** | -0.060 | 0.061* | 0.047 |
| | (0.021) | (0.040) | (0.029) | (0.042) |
| informal childcare | 0.002 | -0.033 | 0.013 | 0.024 |
| | (0.027) | (0.033) | (0.034) | (0.035) |
| lagged formal childcare | | 0.132*** | | |
| | | (0.028) | | |
| lagged informal childcare | | 0.054 | | |
| | | (0.034) | | |
| family learning environment | 0.028** | 0.006 | 0.096*** | 0.051*** |
| | (0.010) | (0.013) | (0.014) | (0.013) |
| lagged family learning environment | | 0.009 | | |
| | | (0.013) | | |
| **Lagged child outcome** | | | | |
| lagged cognitive score | | 0.292*** | | |
| | | (0.024) | | |
| lagged non-cognitive score | | | | 0.539*** |
| | | | | (0.021) |
| Demographic controls [c] | YES | YES | YES | YES |
| Regional FE | YES | YES | YES | YES |
| R-squared | 0.075 | 0.148 | 0.060 | 0.310 |
| Observations | 4589 | 2625 | 4633 | 3094 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level.
[a] The reference group is low-income on-welfare families, and low-income in-work is the treatment group.
[b] The reference group is no non-maternal childcare.
[c] Demographic controls include child gender, birthweight, number of siblings, birth parity, mother's age and age square, dummy for having first degree or above and ethnicity.

Table 1.13: Mechanism: Treatment Effect through Childcare Usage and Family Environment–Second Control Group

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | cognitive | | non-cognitive | |
| | baseline | VA | baseline | VA |
| **Working Status**[a] | | | | |
| low-income in-work | 0.057 | 0.059 | -0.194** | -0.083 |
| | (0.065) | (0.057) | (0.063) | (0.050) |
| **Childcare arrangement** | | | | |
| formal childcare | -0.189 | -0.255 | 0.233 | 0.004 |
| | (0.123) | (0.191) | (0.168) | (0.179) |
| informal childcare | -0.263* | -0.316 | 0.196 | -0.028 |
| | (0.123) | (0.190) | (0.168) | (0.176) |
| family learning environment | 0.044* | 0.034 | 0.103*** | 0.057* |
| | (0.021) | (0.021) | (0.029) | (0.025) |
| **Lagged child outcome** | | | | |
| lagged cognitive score | | 0.244*** | | |
| | | (0.046) | | |
| lagged non-cognitive score | | | | 0.528*** |
| | | | | (0.047) |
| Demographic controls [c] | YES | YES | YES | YES |
| Regional FE | YES | YES | YES | YES |
| R-squared | 0.067 | 0.122 | 0.090 | 0.328 |
| Observations | 895 | 768 | 910 | 786 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level.

[a] The reference group is middle-income in-work families, and low-income in-work is the treatment group.

[b] The reference group is no non-maternal childcare.

[c] Demographic controls include child gender, birthweight, number of siblings, birth parity, mother's age and age square, dummy for having first degree or above and ethnicity.

Table 1.14: Mechanism: Treatment Effect through Income Channel

| | cognitive | | non-cognitive | |
|---|---|---|---|---|
| | baseline | CVA | baseline | VA |
| | (1) | (2) | (3) | (4) |
| **Panel a: Control Group: low-income non-work** | | | | |
| low-income in-work[a] | 0.122 | 0.092 | 0.037 | 0.251 |
| | (0.100) | (0.128) | (0.126) | (0.132) |
| low-work*income | -0.006 | -0.010 | 0.012 | -0.020 |
| | (0.011) | (0.013) | (0.014) | (0.014) |
| income | -0.004 | 0.003 | -0.006 | 0.008 |
| | (0.008) | (0.009) | (0.010) | (0.011) |
| R-squared | 0.067 | 0.134 | 0.049 | 0.316 |
| Observations | 2591 | 1728 | 2600 | 1732 |

| | cognitive | | non-cognitive | |
|---|---|---|---|---|
| | baseline | VA | baseline | VA |
| **Panel b: Control Group: middle-income in-work** | | | | |
| low-income in-work[b] | 0.598 | 0.534 | -0.111 | 0.187 |
| | (0.305) | (0.361) | (0.456) | (0.518) |
| low-work*income | -0.035 | -0.033 | -0.008 | -0.029 |
| | (0.021) | (0.025) | (0.032) | (0.036) |
| income | 0.022 | 0.014 | 0.019 | 0.024 |
| | (0.018) | (0.022) | (0.030) | (0.034) |
| R-squared | 0.067 | 0.118 | 0.102 | 0.347 |
| Observations | 552 | 433 | 559 | 442 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level. All regressions control for demographic characteristics and region fixed effects.

[a] The reference group is low-income on-welfare families, and low-income in-work is the treatment group.

[b] The reference group is middle-income in-work families, and low-income in-work is the treatment group.

Table 1.15: Mechanism: Treatment Effect through Working Hour Channel

| | cognitive | | non-cognitive | |
|---|---|---|---|---|
| | baseline | CVA | baseline | VA |
| | (1) | (2) | (3) | (4) |
| **Panel a: Control Group: low-income non-work** | | | | |
| low-income in-work[a] | 0.058 | 0.022 | 0.157*** | 0.127** |
| | (0.033) | (0.038) | (0.046) | (0.045) |
| low-work*wkhr | -0.013* | -0.007 | -0.014 | -0.006 |
| | (0.006) | (0.008) | (0.008) | (0.011) |
| working hour | 0.013* | 0.006 | 0.014 | 0.004 |
| | (0.006) | (0.008) | (0.008) | (0.011) |
| R-squared | 0.068 | 0.148 | 0.055 | 0.309 |
| Observations | 5533 | 3613 | 5580 | 3655 |
| | cognitive | | non-cognitive | |
| | baseline | VA | baseline | VA |
| **Panel b: Control Group: middle-income in-work** | | | | |
| low-income in-work[b] | -0.301 | -0.044 | -0.135 | 0.160 |
| | (0.234) | (0.255) | (0.203) | (0.220) |
| low-work*wkhr | 0.009 | 0.002 | -0.001 | -0.007 |
| | (0.007) | (0.007) | (0.006) | (0.006) |
| working hour | -0.009 | -0.003 | 0.002 | 0.006 |
| | (0.006) | (0.007) | (0.005) | (0.006) |
| R-squared | 0.060 | 0.125 | 0.088 | 0.297 |
| Observations | 1489 | 1127 | 1510 | 1150 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level. All regressions control for demographic characteristics and region fixed effects.

[a] The reference group is low-income on-welfare families, and low-income in-work is the treatment group.

[b] The reference group is middle-income in-work families, and low-income in-work is the treatment group.

Table 1.16: Robustness Check: the Effect of 30 hour Element of the In-work Benefit Reforms

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | cognitive | | non-cognitive | |
| | baseline | CVA | baseline | VA |
| **Panel a: Control Group: low-income non-work** | | | | |
| low-income working< 30 hours | 0.062** | 0.019 | 0.162*** | 0.088*** |
| | (0.021) | (0.026) | (0.027) | (0.024) |
| low-income working>= 30 hours | 0.048 | -0.020 | 0.128*** | 0.024 |
| | (0.030) | (0.036) | (0.037) | (0.031) |
| lagged low-income working< 30 hours | | 0.079** | | |
| | | (0.027) | | |
| lagged low-income working>= 30 hours | | 0.091* | | |
| | | (0.042) | | |
| lagged cognitive score | | 0.316*** | | |
| | | (0.021) | | |
| lagged non-cognitive score | | | | 0.544*** |
| | | | | (0.020) |
| R-sqr | 0.067 | 0.148 | 0.055 | 0.309 |
| Observations | 5533 | 3613 | 5580 | 3655 |

| | cognitive | | non-cognitive | |
|---|---|---|---|---|
| | baseline | VA | baseline | VA |
| **Panel b: Control Group: middle-income in-work** | | | | |
| low-income working< 30 hours | 0.019 | 0.064 | -0.183*** | -0.075 |
| | (0.051) | (0.048) | (0.055) | (0.046) |
| low-income working>= 30 hours | -0.003 | 0.020 | -0.182** | -0.085 |
| | (0.055) | (0.052) | (0.057) | (0.049) |
| lagged cognitive score | | 0.262*** | | |
| | | (0.038) | | |
| lagged non-cognitive score | | | | 0.496*** |
| | | | | (0.039) |
| R-sqr | 0.059 | 0.126 | 0.088 | 0.297 |
| Observations | 1489 | 1127 | 1510 | 1150 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level. All regressions control for demographic characteristics and region fixed effects.

Table 1.17: Robustness Check: Consistency of WFTC and WTC/CTC reform effects—Difference-in-difference Estimations

| | (1) baseline[1] | (2) cognitive time trend[2] | (3) child FE[3] | (4) baseline | (5) non-cognitive time trend | (6) child FE |
|---|---|---|---|---|---|---|
| treat[a] | 0.100 | 0.079 | 0.009 | 0.109* | 0.093 | 0.034 |
| | (0.054) | (0.058) | (0.065) | (0.055) | (0.059) | (0.069) |
| treat*post | 0.014 | -0.028 | -0.066 | 0.070 | 0.038 | 0.066 |
| | (0.056) | (0.068) | (0.071) | (0.059) | (0.073) | (0.077) |
| post [b] | 0.136*** | 0.150*** | 0.151*** | -0.244*** | -0.234*** | -0.254*** |
| | (0.033) | (0.036) | (0.036) | (0.036) | (0.040) | (0.040) |
| time trend | -0.112*** | -0.120*** | -0.109*** | -0.022* | -0.027 | 0.001 |
| | (0.010) | (0.013) | (0.013) | (0.011) | (0.015) | (0.015) |
| treat*time trend | | 0.020 | 0.023 | | 0.015 | -0.009 |
| | | (0.020) | (0.021) | | (0.022) | (0.021) |
| **Child FE** | NO | NO | YES | NO | NO | YES |
| **Demographic controls** [c] | YES | YES | YES | YES | YES | YES |
| **Regional FE** | YES | YES | YES | YES | YES | YES |
| Observations | 7534 | 7534 | 7534 | 7581 | 7581 | 7581 |
| R-squared | 0.039 | 0.039 | 0.022 | 0.044 | 0.044 | 0.033 |

Note: Standard errors are reported in parentheses and are clustered at child level. *** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level.

[1] This specification refers to equation 1.18.

[2] This specification refers to equation 1.19.

[3] This specification refers to equation 1.20.

[a] Treat refers to the low-income in-work group and the control group is the low-income on-welfare group.

[b] Post refers to post April 2003 period i.e. post WTC reform.

[c] Demographic controls include child gender, birthweight, number of siblings, birth parity, mother's age and age square, dummy for having first degree or above and ethnicity.

# Figures

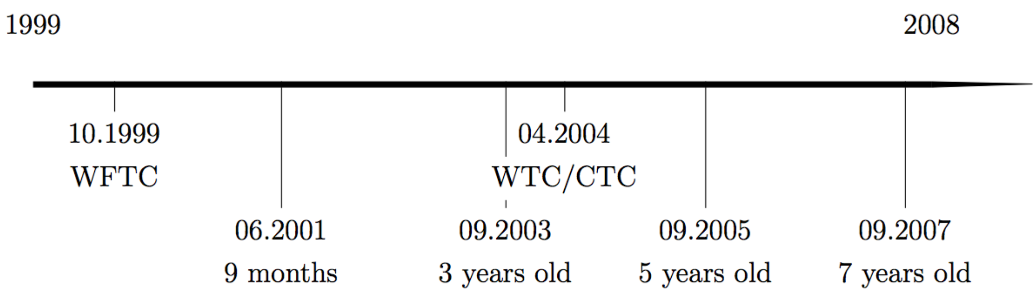
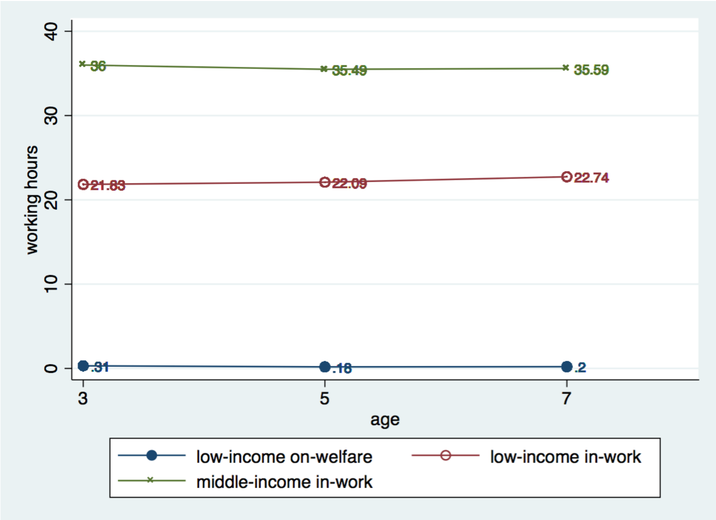



Figure 1.1: Timeline

Note: There are two in-work benefit reforms in the period 1999-2008: the WFTC reform and WTC/CTC reform. The points of observation for children's outcome occurred when they are 9 months, 3, 5 and 7 years old. Due to different interview date, the information documented at wave 2(age 3) could refers to the outcome at before/after the implementation of 2003 WTC/CTC.

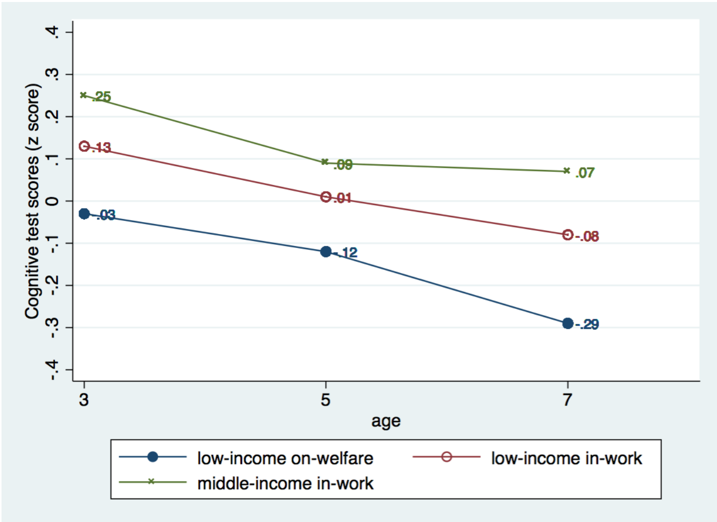

(a) Income gap



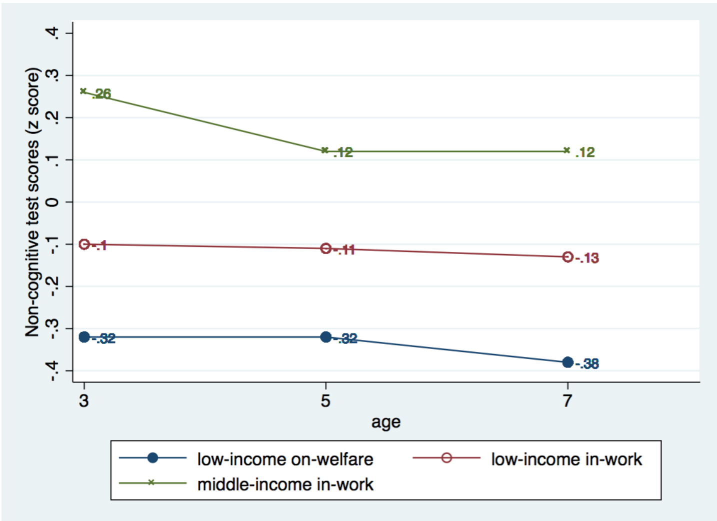

(b) Working hour gap

Figure 1.2: Mother's Working Status by Group

Note: Cells are the average income band or working hour at each survey age. Panel (a) shows the net income gap among low-income on-welfare, low-income in-work and middle-income in-work groups. Panel (b) shows the working hour gap among low-income on-welfare, low-income in-work and middle-income in-work groups.




(a) Cognitive outcome gap

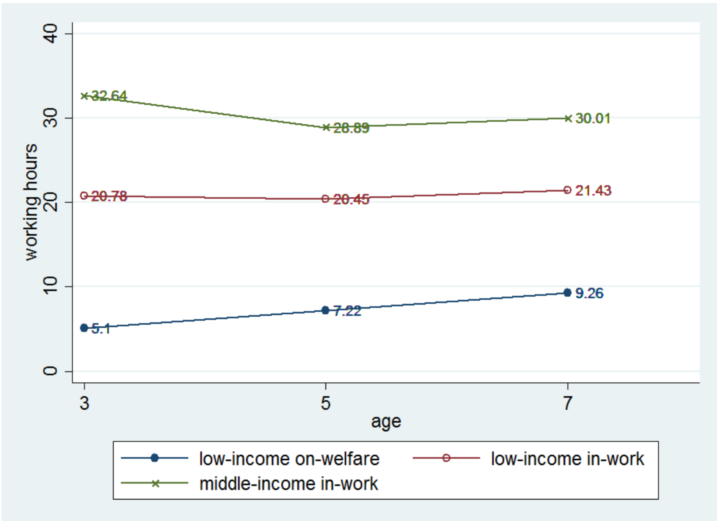



(b) Non-cognitive outcome gap

Figure 1.3: Child Outcome Gap by Group

Note: Cells are the average child outcome at each survey age. Panel (a) shows the cognitive outcome gap among low-income on-welfare, low-income in-work and middle-income in-work groups. Panel (b) shows the non-cognitive outcome gap among low-income on-welfare, low-income in-work and middle-income in-work groups.

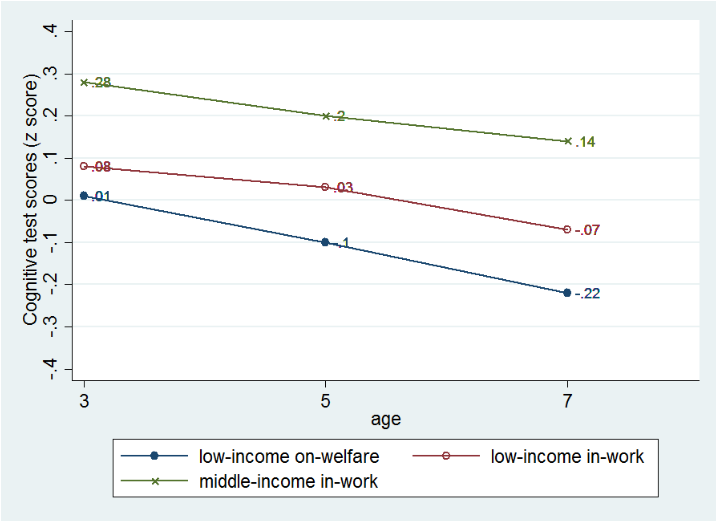

(a) Income gap



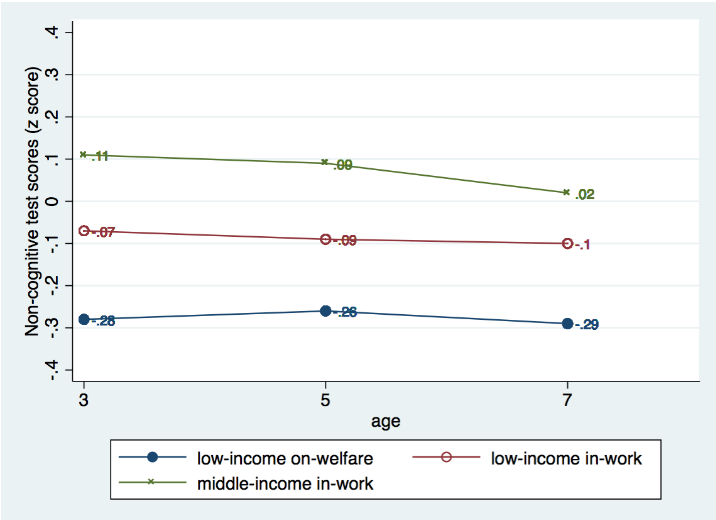

(b) Working hour gap

Figure 1.4: Robustness Check: Alternative Definition of Groups—Mother's Working Status by Group

Note: Cells are the average income band or working hour by group and survey age. The composition of each group are fixed under the WFTC regime, i.e. the definition of treatment and control group is defined based on the WFTC rule. Panel (a) shows the net income gap among low-income on-welfare, low-income in-work and middle-income in-work groups. Panel (b) shows the working hour gap among low-income on-welfare, low-income in-work and middle-income in-work groups.




(a) Cognitive outcome gap




(b) Non-cognitive outcome gap

Figure 1.5: Robustness Check: Alternative Definition of Groups—Child Outcome Gap by Group

Note: Cells are the average child outcome at each survey age. The composition of each group are fixed under the WFTC regime, i.e. the definition of treatment and control group is defined based on the WFTC rule. Panel (a) shows the cognitive outcome gap among low-income on-welfare, low-income in-work and middle-income in-work groups. Panel (b) shows the non-cognitive outcome gap among low-income on-welfare, low-income in-work and middle-income in-work groups.

# Chapter 2

## Is the Quantity-quality Trade-off Real?

## Quasi-experimental Evidence from China

## Abstract

China has implemented several family planning policies during the last four decades. For the purpose of understanding the implication and consequences of controlling population size, this paper investigates the relationship between fertility and children's educational outcome(quantity-quality(Q-Q) trade-off). It departs from previous literature in accounting for the non-linear distribution both within each birth parities and across birth parities. Specifically, the non-normal distribution within each birth parity is accounted for by applying Generalized Method of Moments(GMM). It respects the limited dependent nature of both outcome and endogenous variables. In contrast with conventional 2SLS estimates, GMM estimates give significantly negative effects of fertility on education outcome of children, which is consistent with the Becker-Lewis model that an increase in child quantity increases the shadow price of quality. Moreover, the Q-Q trade-off non-linearly decreases with family size and shows heterogeneous effects by birth order. The identification strategy exploits variation in family size that is induced by twin births and first child's gender. This paper is the first to apply first child's gender as an instrumental variable to explore the Q-Q trade-off in the context of China where son preferences are prevailing. This instrument offers the possibility to test the Q-Q trade-off at a lower fertility margin than the typical instruments based on twins and same-sex composition.

**Key Words**: Fertility, education, quality-quantity trade-off, GMM, twin births, first child's gender
**JEL Classification:**: C26, J13, J18, J24, O15, P20

## 2.1 Introduction

Under the general assumption that parents invest in all their children equally, Becker and Nigel (1973) propose that an exogenous increase in the number of children increases the marginal cost of investment in their quality. Motivated by the theoretical model of the quantity-quality(Q-Q) trade-off (Becker and Nigel, 1973, 1976; Willis, 1973), the relationship between family size and child outcome has been one of the most enduring in economics. On the one hand, the declining quantity of children will free up resources for human capital investment; on the other hand, larger family size may benefit from the economies of scale and the social interaction of children, as Black et al. (2005) argued, children with fewer siblings may not be better off than if they are in a larger family.

If the Q-Q trade-off exists, it suggests that parents who have less children allocate more of their time and resources to each of the children assuming that parents invest in children evenly and holding income constant. Furthermore, the trade-off implies that those policies aiming at curbing population growth or subsidizing families with fewer children may contribute to better health, educational and behavioural outcomes of each child. In turn, it may reduce abortion rates, crime rates and poverty rates in the society (Schultz, 2007). For example, if couples perceive this Q-Q trade-off story and prefer better quality children, parents could be motivated to avoid unwanted births and reduce social problems. Therefore, a significant trade-off provides evidence for government to advance contraceptive knowledge and subsidize birth control.

However, population control policy is unsustainable for long term economic growth, especially for ageing societies such as some western countries and China. China implemented One-Child Policy(OCP) in 1978 to curb the rapid population growth, but it realized recently that the OCP gave substantial pressure for the only child to give elderly care for parents and for the society to bear the cost of ageing population structure. Thus the OCP was relaxed to some extent in 2014(selective two-child policy), and two-child policy has been implemented since early 2016. These shifts in family planning policy induces complicated demographic transition, thus it would be important to understand the relationship between fertility and human capital accumulation. If the Q-Q trade-off exists, policies aiming at maintaining sustainable economic growth should be equipped with policies that alleviate the adverse effect of fertility, such as more generous maternal leave, public education and health support for the family. It is also important to test the existence of a cutoff point for Q-Q trade-off. If non monotonicity and heterogeneity exist in quantity effects, then a trade-off may be found in a certain range of family size and households with certain characteristics. Thus policies should be adjusted accordingly.

While the theoretical model implies a negative trade-off, empirical results of the Q-Q trade-off need to be treated critically. It is well established that there is no Q-Q trade-off in developed countries, but developing societies tend to have less conclusive findings. Due to the higher quality of public education system and better government support for childbearing and childcare, it is likely that the Q-Q trade-off is less evident in more developed economies. This study focuses on children born in 1970s with a majority of them living in rural China, they suffered from poor education system and were born before[1] the implementation of One Child Policy.

The compulsory schooling law in China was not implemented until 1986. It mandates six-year primary school and three-year junior high school. The primary school enrolment rate was quite low at the beginning of Chinese economic reform in 1978. Only 36.6 percent of the counties has widespread access to primary school education for schooling age children in 1985, and the enrolment rate of primary school and junior high school was 95.9 percent and 68.4 percent, respectively[2]. The implementation of compulsory schooling was not uniform across the country. It largely depended on local government budget which made the access to education more difficult in poor rural areas. In poor areas, even if public school was provided, it was not totally free, parents still had to pay the tuition and other fees[3]. Thus the school enrolment of children were subject to the choices of parents and local government budget and facilities prior to 1986. The sample under study in this paper is composed of children born in a period which is before the implementation of compulsory schooling law.

This paper presents new evidence on the child quantity(sibling size)-quality(schooling) trade-off in the context of China. The identification strategy exploits variation in family size that is induced by the gender of first birth and twin births. Typically previous literature tends to estimate the trade-off using linear probability approach and only considers non-linearity across birth parities[4]. This paper departs from previous literature in accounting for the non-linear distribution both within each birth parities and across

[1]The sample also includes children born in early 1980s, which may result in the estimates being contaminated by the OCP. However, robustness check based on an earlier period sample(1982 census) give similar result.

[2]source from the website of the National People's Congress of the People's Republic of China `http://www.npc.gov.cn/npc/xinwen/rdlt/fzjs/2009-02/13/content_1470214.htm`

[3]It is not until 2000 that the coverage of nine-year compulsory schooling reached 85% and the illiteracy rate reduced to less than 5%.

[4]Here birth parities means at least 1 birth sample(families with at least 1 birth), at least 2 births sample(families with at least 2 births) and at least 3 births sample(families with at least 3 births).

birth parities. Specifically, the non-normal distribution within each birth parity is accounted for by applying Generalized Method of Moments(GMM). It respects the binary nature of the outcome variables and the count nature of the endogenous variable, thus it gives a consistent and more efficient estimator of the Q-Q trade-off. The non-linear distribution of fertility effects across birth parities is accounted for by estimating Q-Q trade-off at each birth parity separately(Black et al., 2005; Angrist et al., 2010).

This paper is the first to apply the variation in fertility generated by first child's gender into testing the Q-Q trade-off in the context of China where son preferences are prevailing. This is largely influenced by the Confucianism that it stresses the importance of patrilineality within Chinese society (Das Gupta et al., 2003). Patrilineality enforces sons to continue the family lineage and inheritance. It is also widely believed that girls marry "out" of their natal families and parents cannot rely on girls for old-age support. Therefore, families with son preferences will generally have larger family size if the first birth is a girl than a boy. This instrument uncovers a channel to test the Q-Q trade-off of a marginal second birth on the first born, while typically twins and parental preference for a mixed gender composition estimate the effect of a marginal third or higher birth order child on older siblings[1]. Moreover, using it as an instrumental variable provides a sensitivity test on twinning as an instrument.

I find significantly adverse effects of one additional birth on the education outcome of the firstborn child for families with at least one or two births, regardless of the instruments. The Q-Q trade-off is estimated separately for firstborn and secondborn children of families with three or more children, as it is reasonable to believe that children of different birth order would present different response to an additional birth. Indeed, Q-Q trade-off is bigger for the second-born children than for the first-born children. Generally the Q-Q trade-off is more evident in a relatively smaller family. It maybe due to the fact that conditional on a bigger family size, the resource is spread more thinly and thus the marginal decrease in the welfare of older sibling declines with child quantity. It provides some evidence on the non-linearity of trade-off as family size goes up.

Although exploring potential mechanisms through which family size affects children's outcome is beyond the scope of this paper, I explored mother's labour supply adjustment and economies of scale channels. Basically, mother's labour supply stayed

[1]In this paper marginal child refers to the child that is born due to the instrument, for example, a marginal third birth refers to the additional child induced by twins at the second birth.

stable when family size increased from 2 to 3 but dropped significantly when fertility raised from 3 to 4. It has been documented that maternal time is very important in the early stage of the children's development. If reduction in labour supply indicates an increase in maternal supervision time with their children, then this labour supply adjustment could possibly explain the finding that 2+ families showed a significant Q-Q trade-off while 3+ families displayed no Q-Q trade-off. The evidence on economies of scale in terms of sharing schooling costs was mixed, but in general it favoured the channel that children from a bigger family enjoyed economies of scale[1].

Several tests are performed to test the validity of instrumental variables since twin births and girl at first birth may affect the schooling outcome independently from family size. Twinning may affect the outcome of lower parity children through zero spacing and lower birthweight. I test the existence of these mechanisms, and the results suggest that estimators of Q-Q effects are probably positively biased. Thus the true adverse effects of family size on the schooling performance of the children are bigger.

The validity of firstborn gender implies that it should be random so that it is not correlated to any parental preference. Thus it requires assumptions to be made on parental investment in education, i.e. parents have no gender-biased investment preference and child's ability is independent of gender. This paper tests this hypothesis in the following ways.

First, if the sex ratio at the first birth is within the natural range[2], then it can at least rule out pre-natal sex preferences. One would worry about the technology advancement in invalidating the gender-related instruments. Ultrasound diagnosis of fetal gender became available since early 1980s, but sex-selective abortion techniques became widely available only after 1986/7 (Lin et al., 2008). My sample period covers children born between 1974 and 1984, among which the younger cohorts could be affected by theses technologies but only to a weak extent. A simple way to check sex-selective abortion on the firstborn is to compare mother's age at first birth with boy versus girl. The idea is that if abortion exists, it should delay son's birth. However, I find no significant difference between mother's age at firstborn boy and at firstborn girl. Additionally, the infant mortality rate was 70 deaths per 1000 births, among which boy death was higher than girl death. Furthermore, to rule out the contamination of technology effect,

[1]Results on these mechanisms are not attached in this version of paper, but they can be provided upon request.

[2]The World Health Organization(WHO) defines natural sex ratio at birth to be 105 males for every 100 females. More information at http://www.searo.who.int/entity/health_situation_trends/data/chi/sex-ratio/en/

results are also provided using 1982 census which is consists of children born before the availability of these technologies. Second, possibly the validity can be tested by comparing the educational outcomes of children for different gender holding family size constant. If families with more boys have better outcomes than those with more girls, conditional on family size, then it indicates a post-natal sex-selective investment, assuming child ability is independent of gender. This paper find no evidence of post-natal gender biased investment behaviour from this perspective. Third, if gender at first birth does not affect child outcome independently from family size, then one would expect no reduced-form effects in a sample where the first child gender should matter the least. Unfortunately, this validity test fails to rule out the direct impact of first child's gender, but the direction of bias indicates that the true Q-Q tradeoff is smaller.

Previous twins-based studies mainly examine the effects of family size on lower birth order children, the effects on outcome of twins themselves are often overlooked. Rosenzweig and Zhang (2009) suggest that the effect of twinning on twins and non-twins provides upper and lower bound estimation of Q-Q trade-off, under the assumption that parents allocate resources towards non-twins due to the birthweight deficit. Therefore, this study also employs twin at first birth as instrument to bound the traditional twin estimators. Applying their estimation strategy, I find support of parental reinforcement behaviour, implying estimators of lower bound are positively biased. Thus, the true Q-Q trade-off based on twin instrument is bigger than what is showed in this study.

There is evidence of heterogeneity in supporting that the trade-off is mainly driven by children of rural households and of less educated mothers. Possible explanation could be that children from relatively disadvantaged background experienced less developed education system and government support, and were more likely to be financially constrained due to the less developed capital market.

The paper unfolds as follows: section 2.2 presents the institutional background and literature review on the analysis of quantity-quality trade-off. Section 2.3 provides the estimation strategy. Section 2.4 describes the data source and descriptive statistics. Section 2.5 provides empirical results and robustness checks. Finally, section 2.6 draws conclusion.

## 2.2 Institutional Background and Literature Review

### 2.2.1 Policy background

After two decades of rapid population growth, the Chinese central government enacted a series of birth planning campaign to curb population growth in 1971. At first, the main objective was to encourage people to get married later, to expand birth spacing between births and to have fewer children. Later, a stricter policy – the One Child Policy – was announced in 1978 and tightly enforced across the country in 1980. Each household was allowed to have only one child, a second child was permitted only under extreme circumstances. However, it only applied to individuals of Han ethnicity, ethnic minorities were exempt from this policy. It generates additional variation in family size across ethnicities.

Considering the difficulties in enforcing the policy and the high labour demand in rural areas, this restriction was relaxed in the manner that rural couples were allowed to have a second birth after a specified interval if the first child was a girl in 1984. However, in some rural areas, all couples were allowed to have two children. Violators of the family planning policy will mainly be punished in the form of monetary penalties. The amount of fines are generally big and varies with regions. Under certain circumstance, people can even lose their jobs or permanently lose the chance to get promoted, for instance if one of the couple works in the government or state-owned enterprise.

There has been a gradual relaxation of ChinaâĂŹs family planning policies in recent years. The selective Two-Child policy was implemented in 2014. Couples in many parts of the country have been allowed to have two children if one parent was an only child. However, eligible couples did not illustrate strong wishes of giving additional birth. Thus this policy was far from reaching the government goal. In early 2016, Two-Child policy has been enacted to allow for maximum 2 children in all families.

### 2.2.2 Literature Review

A main implication of the quantity-quality trade-off model (Becker and Nigel, 1973, 1976) is that the marginal cost of quality increases with quantity, holding the income constant. However, empirical research, to some extent, diverges from the theory implication.

Rosenzweig and Wolpin (1980) are the first to confirm the hypothesis of Becker and Nigel (1973, 1976) by using twins as an exogenous increase in fertility in India. However, their estimation is imprecise due to the fact that the sample size is small(25 twins pairs), and the selection bias problem resulting from using outcomes of all children. Moreover, the internal validity of twinning as an instrumental variable relies on a large sample data which helps to mop up the infrequency of multiple births.

More recent studies based on developed societies tends to find negligible effects of family size on educational attainment of children. Using a data set on the entire population of Norway, Black et al. (2005) argue that previous work on the impacts of family size[1] would be biased if not controlling for birth order effects. In fact, they find negligible family size impacts on children's education when controlling for birth order. Instead of affecting the quality of each child, family size may only affect the marginal child through the impact of birth order. In particular, higher birth order has significant negative effects on child's education, adult earnings, employment and teenage childbearing. Theoretically, there are some models predicting birth order effects, say the optimal stopping model(continue to have children until they have a poor quality child). However, later research fails to find this pattern of birth order effects. One candidate explanation is that the Q-Q trade-off is offset by the good education system and generous welfare in well-off countries.

As pointed out by Imbens and Angrist (1994), the estimations generated by any particular IV approach capture only the effects on individuals affected by that instrument, which should be interpreted as the local average treatment effect (LATE). Angrist et al. (2010) combine evidence from multiple sources of variation in family size by separately and jointly using the incidence of twin births and same-sex sibling pairs. This method captures the effect of fertility across different treatment groups and therefore increases the range of variation. Their results show no evidence of Q-Q trade-off in Israel. However, their study fails to capture the effects on the marginal child, and the estimated effects on college attendance show large standard error. Additionally, a limitation of using twins at second/third birth and preference for a mixed gender composition of children is that they give only the marginal effects of a third/fourth child, while the marginal effects of a first or second child is technically inaccessible.

Developing countries differ from developed countries in many aspects, such as economy status, public education system, government support for childbearing and

[1] This paper uses fertility effect, family size effect, sibling size effect and quantity effects interchangeably. They all refers to the effect of number of children.

childcare, and social norms, which may render the trade-off more evident in the developing countries.

Lee (2008) studies the trade-off by exploiting variation in family size under son preferences in South Korea. He proposes an new instrumental variable – the first child's gender and directly tests the effect by using measurements of parental monetary investment in children's education, as he argues that the economics theory of Q-Q trade-off is about parental choices. He finds that per-child investment is reduced by 25.4% for children with one sibling, and by 42.4% for children with two siblings compare to the only child. The trade-off goes up as the family size is bigger. However, using parental monetary investment solely is not enough to capture the whole picture of Q-Q trade-off. For example, parental supervision time has been shown in economic and sociology literature that plays an important role for children's human capital accumulation, especially at the early stage.

Following the method of Black et al. (2005), Li et al. (2008) estimate the trade-off in China using 1990 Chinese Population Census, they find no effect of family size on first-born's education by exploiting the exogenous variation in family size induced by twins at second birth, and a significant trade-off on average educational outcome of first two births by using twins at third birth. They show that controlling for birth order do not alter the Q-Q trade-off effects, which contradicts the finding of Black et al. (2005). However, they treat an ordered discrete dependent variable using ordinary least square(OLS) estimation, which renders their results less convincing. Additionally, they do not estimate fertility effect for the first and second born children separately, in the at least three births family.

Rosenzweig and Zhang (2009) use a new data set, the Chinese Child Twins Survey(CCTS), to quantify the Q-Q trade-off of children. They also find a significantly negative relationship between family size and child quality. They exploit the effects by using the incidence of twin births that for the first time taking into account the effects of birth-weight deficit and close spacing of twins. Their study firstly proposes that the impacts of twinning on older non-twin births and on twin births themselves provide the lower and upper bounds of true impacts of the family size on average child quality. Moreover, the lower bound is positively biased if parents exhibit reinforcement behaviour. Thus true Q-Q trade-off should be bigger than what have been found in previous literature, if not accounting for variations in endowment. However, their strategy ignores the fact that birthweight itself is endogenous because it is affected by twinning. In addition, their approach is more likely to be a reduced form estimation by using twins, instead of using twins as an instrument to solve the endogeneity problem.

Based on different datasets in China, Li et al. (2008) find no evidence of Q-Q trade-off, but Rosenzweig and Zhang (2009) show negative effects of fertility rates on education outcomes of children by exploiting variation in family size using twins. Instead of the twins IV, China's family planning policy provides a good opportunity for researchers to investigate the quantity-quality trade-off. Using one child policy as an instrument provides a different LATE, Qian (2009) finds positive family size effects and Liu (2014) shows no significant effects on education outcome. The heterogeneity in the Q-Q trade-off suggests that it is sensitive to the sample period and composition.

Qian (2009) makes a breakthrough by using the relaxation of China's one child policy in rural China to instrument fertility. The sample size is substantially boosted by matching the 1989 CHNS data with the 1990 1% sample census at the county level. In particular, she exploits the exogenous variation by a triple interaction of the child's gender, year of birth and region of birth. Results show that an extra child increases school enrolment of first-born child by 16% for one-child family. She proposes several mechanisms for explaining the positive correlation and finds support in economies of scale in schooling cost and increasing mother labour supply for rising demand on cash for children's future education cost. However, the variation in the implementation of the One Child Policy is likely to correlate with the demand of children, thus the fertility effects derived from the policy are likely to be negatively biased.

In order to partly address the potential correlation problem, Liu (2014) uses two community level variables to control for the potential correlation between parental preference and local family planning polices. His study employs three instruments for fertility: fines for unsanctioned births, the eligibility for having two children, and their interactions. Results show weak fertility effects on children's educational attainment. He suggests that government interventions in education partly offset the effects of parental investments in children's schooling, resulting in a weak Q-Q trade-off.

Prior empirical studies that explore exogenous variations in family size due to twin births and sex composition mostly impose a constant fertility effects on child outcome, Mogstad and Wiswall (2016) re-examine this relationship by allowing an unrestricted functional form. They propose a non-parametric estimation that allows identification of marginal fertility effects, in comparison to the total fertility effects studied previously (Black et al., 2005). They find that previous conclusion of no Q-Q trade-off is an artifact of the linear specification in fertility, which masks significant marginal fertility effects. They conclude a Q-Q tradeoff in large families and Q-Q complementary in small families.

This paper departs from previous empirical work in accounting for the non-normal distribution of outcome and endogenous variables within each birth parity, rather than only across parities. In addition, it explores an additional dimension of marginal fertility effect.

## 2.3 Methodology

### 2.3.1 Instrumental Variables

Child quantity and quality are affected by parental preferences, household characteristics and child's innate ability. It is not possible to fully observe all the factors that affect child quality, hence child quality is endogenously determined by quantity. In order to identify a causal relationship, I use the gender of firstborn and twin births as instrumental variables to capture potentially exogenous variations in family size.

There are three main samples in my estimation, at least one birth, at least two births and at least three births sample. This analysis is limited to lower birth order (*N-1* births) child(ren) in at least *N* births sample as they are the only unconditional treatment units in the household. In specific, at least one birth sample(1+) consists of firstborns from families with one or more births, and the family size is instrumented by the gender of first birth. At least two births sample(2+) consists of firstborns from families with two or more births, and sibling size is instrumented by twins at second birth. At least three births sample(3+) consists of firstborns and secondborns from families with three or more births, fertility is instrumented by twinning at third birth. These three samples are further divided by outcome variables, as it will be explained in section 2.4, due to different exposed age groups.

This paper is the first one to apply the gender of first birth to test Q-Q trade-off in the context of China where traditionally son preferences is influenced by the Confuciusim. The argument for relevance of this instrument is that families with son preferences will generally have larger family size if the first birth is a girl than a boy. Moreover, given the biological time constraint for childbearing, families with a first born girl will shorten the birth spacing if they try to have a boy sooner. By the same logic, a girl at first birth is also a good predictor for a third birth. For example, if families have first two girl births and prefer to have a son, they will try to have an additional birth and should do so more quickly. In general, a girl at first birth is likely to contribute to a higher fertility rate and shorter birth spacing. Lee (2008) proposes two relevant tests of first child's sex as an instrumental variable for fertility.

The first test estimates a parity progression model under the hypothesis that parents having daughters in lower birth orders should end up with larger family sizes if there were son preferences. This model requires that households' decisions about future childbearing are fully observed, so that it will be tested only among those who have completed fertility. The sample is confined to mothers aged above 40 as conventional demographic paper assumes that this is the age range that women are no longer fertile. I estimate a logit model where the main variable of interest is the sex composition of siblings and the dependent variable is an indicator for whether the family has more than $N$ child/children birth. Mother's age at first birth, current age, birth year, education, ethnicity, geographic location of parents are controlled for. The first two columns of table 2.1 show that a girl at first birth increases the probability of having a second child by 5.5% and the likelihood to have a third child by 9.5%. The second column also shows that having same-sex children at first two births increases the probability to have a third child by 11.2% while the third column suggests that the increase in family size is mainly due to two girl births (9.4%). Having two sequential boy births does not affect the family size.

The other model for testing son preferences is a hazard model of fertility timing. The hypothesis in this model is that families should advance the timing of having a second child if the first birth is a girl than a boy given the time constraint of childbearing biologically. Similarly, the birth spacing between the second and third child should be shortened following two consecutive girl births. I estimate a Weibull hazard model where the key variables of interests are sex compositions of siblings and the outcome variable is the birth spacing between corresponding births. To avoid bias that caused by changes of time-variant variables during the interval, only time-invariant variables are included as controls. Right panel of table 2.1 reports relative hazard ratios. The fourth column shows that the hazard rate of having a second child increased 56% by a first girl birth. As before in the parity progression model, the positive effect of samesex sibling in inducing an additional birth is mainly resulted from two consecutive girl births(95%). Two consecutive son births decrease the hazard rate of having a third child by 26%. These results are consistent with the hypothesis that an initial girl birth(births) advances the timing of the second(third) child while boy births delay the additional childbearing.

The rational for twins instrument is that twin births at any parity is random and the subjects of interest are older siblings. The internal validity of twins instrument relies on its randomness conditional on some observed biological factors. One concern would invalidate this instrument is the use of fertility drugs that boosts the chance of having multiple births. However, the introduction of this drug started in late 1980s, which is unlikely to affect my sample. In addition, several robustness checks are carried out in this study with regard to the closer spacing, lower birthweight, economies of scale issues of using twinning as an instrument.

Angrist et al. (2010) point out that combining evidence from multiple sources of variation is important since different instruments are potentially subject to different omitted variables biases. For example, the occurrence of twins varies with maternal age, and twin births may affect outcome directly through birth spacing. Although first child's sex instrument is not suffering from those biases, it may directly affect outcome through gender specific characteristics and postnatal investment preferences of parents. Therefore, estimations based on different instruments provide specification check on themselves.

Table 2.2 shows strong first stage effects. A girl at first birth increases fertility by about 28% to 30% for 1+ households, and twin births increase the family size by about 37% to 51% for 1+ sample, by 56% for 2+ sample, and by 42% to 55% for 3+ sample, depending on the exposed subsamples and birth orders. Figure 2.1 plots first stage estimates of the effect of twins and first child's sex on fertility. Twinning captures shifts in family size over a narrow range that is close to the parity of twins birth, while the first child's sex instrument captures a relatively wider range of fertility variation. For example, the effect of twins at second birth increases the fertility from 2 to 3 with no further effects on higher birth parities. While those families who have son preferences will tend to have children until obtaining a son, and significance of the effects fades out when family size go beyond 5.

### 2.3.2 Rationale for Using Non-linear Model

Previous work mainly uses conventional 2SLS to estimate the fertility effects, while typically the normal distribution assumptions for outcome and endogenous variables are violated. The conventional 2SLS approach can be characterized as follows (Black et al., 2005; Angrist et al., 2010):

$$N_i = Z_i\alpha + X_i\beta + e_i \tag{2.1}$$

$$Y_i = \gamma N_i + X_i\theta + v_i \tag{2.2}$$

where $N_i$ is the number of children in the family(quantity), $Z_i$ stands for instruments, it's a dummy for the gender of the firstborn at 1+ sample, or twin births at parity $N$ for at least $N$ births sample, $X_i$ is a set of control variables, $Y_i$ is a measure of child's educational outcome(quality).

#### 2.3.2.1 Graphic Evidence

Fertility, an endogenous variable in Q-Q trade-off literature, has traditionally been estimated using Ordinary Least Squares(OLS) with normal error terms. Since fertility is a discrete count variable, the normal distribution assumption is inappropriate. More importantly, the assumption of constant variance is invalid. Figure 2.2 explores the form of mean-variance association for households with at least one child. To make the evidence graphically representative, the moments are computed conditional on the education level of mother and mother's age at giving first birth. Clearly, the variance is not constant. Although the variance is not precisely equal to the mean, the plot suggests a proportional relationship between them. Thus it can do more justice to the data by using Poisson estimations than by relying on OLS (Rodriguez, 2007).

Recent literature has realized the non-linear fertility effects and anaylized the Q-Q trade-off for different family size, separately. However, they only consider the non-linearity across parities, but the non-normal distribution within parity is overlooked. Specifically, this paper argues for a GMM approach to account for the count nature of the endogeneous variable and binary nature of the outcome variables. This analysis shows that the use of this approach makes a significant difference in estimating the fertility effects. Following the visual assessment strategy proposed by Silva et al. (2014), figure 2.3 provides visual comparisons of the traditional approach and the proposed method on model fitness[1]. For ease of demonstration, equation 2.1 and 2.2 can be rewritten as the following,

$$N_i = V_i\sigma + e_i \tag{2.3}$$

[1] Bear in mind that the true estimation process is not constructed in two stages, but showing the estimation step by step helps to provide clear rationale for using non-linear approach.

$$Y_i = W_i\phi + v_i \tag{2.4}$$

where $V_i$ includes all the covariates in the first stage of equation 2.1 and $W_i$ includes all the covariates in the second stage of 2.2.

Figure 2.3 focus on families with at least one birth, examining the effects of an additional birth on the probability of completing primary school for the first-born children. The instrument used is the gender of first birth. Panel (a) and (b) of figure 2.3 estimate the first stage and compare the fitness of OLS and Poisson regression. Plots show the value of $N_i$ and parametric fits of $E[N_i|V_i]$ versus the estimated single index $V_i\hat{\sigma}$. In addition, nonparametric fits are displayed to provide a robustness check of the parametric fit, in the sense that parametric one proposes a model to fit the data while nonparametric one let the data commands the regression. Here the nonparametric fit uses a kernel regression of $N_i$ on the fitted value obtained above. Panel (c) and (d) of figure 2.3 estimate the second stage and compare the fitness of OLS and Logit model. They show plots of $Y_i$ and of parametric fits of $E[Y_i|W_i]$ versus the estimated single index $W_i\hat{\phi}$. Similarly nonparametric fits are added to plots.

All of the OLS estimators have the expected sign but the magnitude of marginal effects are clearly erroneous. For example, figure 2.3(c) shows that the fitted value of $E[Y_i|W_i]$ goes beyond zero and way below the upper bound of probability one. As a result, the parametric and non-parametric fits largely diverge from each other, which indicates a mismeasurement of partial effects for most of the observations. The departure of OLS fit from kernel fit may explain the failure of proving a Q-Q trade-off previously. In contrast, figure 2.3(d) displays an overlapping, to a large extent, of Logit and kernel regression, which shows a large improvement on the model fitting. The slight diverge of two regression fits at the upper tail is likely due to the fact that fewer children are enrolled at junior high school than those who are not.

Similar analysis is applied in figure 2.4 where the outcome variable is the probability of enrolling in Junior high school with twins at the second birth as an instrument variable[1]. Non-linear fitting shows better performance at the distribution boundaries. While the performance of non-linear fitting is not necessarily better than OLS within the boundaries, the non-linear estimation at least provides another bound of the Q-Q trade-off.

To sum up, the merit of using non-linear specification are manifold. First, it allows

[1]Here only two of the many regression are showed graphically, more plots are provided on request.

for non-normal distribution of the error term. Second, it relaxes the constant variance assumption imposed on Ordinary Least Square estimation. Third, it captures the discrete nature of the data, and allows inference to be analysed on the probability of event occurrence. Although coefficients can be estimated consistently by OLS, the inference based on the estimated covariance matrix is not valid.

Figure 2.3 and 2.4 have shown that non-linear fitting approximates data better than ordinary least squares at the distribution boundaries. This is of particular importance if the true value of outcome variable have higher probability locating closer to 0 or 1. The econometric estimation will be presented as follows.

#### 2.3.2.2 Econometrics Reasoning

Equation 2.5 uses a Poisson regression to estimate first stage, but directly plugging in the first-stage fitted value to a Logit estimation—equation 2.6 would fall into the forbidden regression problem (Angrist and Pischke, 2008).

$$E(N_i|X_i, Z_i) = exp(Z_i\alpha + X_i\beta) \tag{2.5}$$

$$Y_i = F(\gamma N_i + X_i\theta) + \varepsilon_i \tag{2.6}$$

Estimation in stages, by replacing $N$ with its estimated conditional mean, does not give consistent estimates of the parameters (Windmeijer and Santos Silva, 1997; Wooldridge, 2010).

$$Y_i = F(\gamma exp(Z_i\alpha + X_i\beta) + X_i\theta + \gamma\eta_i) + \varepsilon_i \tag{2.7}$$

where

$$\eta_i = N_i - exp(Z_i\alpha + X_i\beta) \tag{2.8}$$

Equation 2.9 shows that the moments of $\eta_i$ depend on parameters and regressors.

$$E(\eta_i^2|X_i, Z_i) = exp(Z_i\alpha + X_i\beta) \tag{2.9}$$

Angrist and Pischke (2008) propose a simple alternative to the forbidden second step by using the nonlinear fitted values as instruments, instead of directly plugging in nonlinear fitted values. It has been shown graphically that nonlinear estimators give better approximations to the first stage conditional expectation function than the linear model, thus an additional advantage of non-linear modelling is the improvement in efficiency.

Furthermore, as suggested by Windmeijer and Santos Silva (1997) and Wooldridge (2010), a consistent estimator would be the Generalized Method of Moments(GMM) estimator, and the nonlinear fitted values are used as instruments. Following Windmeijer and Santos Silva (1997), a GMM approach with $exp(Z_i\hat{\alpha} + X_i\hat{\beta})$ as an instrument has been applied to this analysis. Using the fitted values, obtained from the Poisson regression, as instruments provide a best approximation to the optimal instrument set. Figure 2.5 and 2.6 show the comparison of 2SLS and GMM estimator, clearly the advantage of GMM estimation exhibits at the boundaries. Thus the GMM estimation is appropriate for this study in the sense that outcome variables are more located at the upper tail of the distribution.

Additionally, number of children in a household are likely to display under-dispersion property, an ideal model to control for dispersion (Winkelmann and Zimmermann, 1994; Wang and Famoye, 1997) would be Generalized Poisson regression. However, this attempt to improve the instrument will have only a marginal impact on the final estimator, but will result in more of computational complication due to the additional parameters to be estimated. Thus, this paper uses Poisson regression to obtain the instrument.

## 2.4 Data

The dataset is the 1% sample of 1990 Chinese Population Census, it covers 11,835,947 individuals from 3,260,243 households. The 1990 census questionnaire contains detailed demographic information, including the following: name, relationship, sex, age, nationality, registration status, education, industry, occupation, unemployment status, marital status, residence in 1985, number of children ever born, number of children surviving, birth order in 1989 etc. The census provides no direct information on parental-children linkage, the relationship is thus inferred using the census question on what is each household member's relationship to the head and using the IPUMS matching rule[1].

For the purpose of this analysis, information on both parents and children are needed. However, the drawback of using census data is that it has no record on individual who left the family. Thereby this analysis requires households to have

[1] Information on the IPUMS algorithm is available at http://www.ipums.umn.edu

parents and children co-residence in the household[1], which drops 33% of the households. This reduction in sample size is largely due to elderly/very young families with no children presenting or families with no parents presenting. This analysis also requires observations of all alive children in the family since the test of Q-Q trade-off is built on resource competition or sharing. Thus it excludes families with reported number of children alive unequal to the constructed family size. There are 29% (690,924) of the households with reported number of children alive bigger than constructed family size. It suggests that those families having non-coresident children. Additionally 1% of the households have reported number of children smaller than the constructed family size, which could be attributed to measurement error or misreporting.

This study further restricts the sample by excluding mothers giving birth before age 16(0.2% of the sample); excluding families that migrated in the last 5 years[2]; excluding families with multiple birth (except twins birth) since they have closer spacing and endowment deficit compared to the singleton child(125 households are dropped). Finally, the sample size is 1,656,679 households.

This analysis focuses on three domains of outcome[3]: the probability of primary school enrolment, primary school graduation and junior high school enrolment. School enrolment is a binary variable that equals to one if a child was enrolled in school(attending) or dropped out or graduated, and zero if a child had never enrolled school. Graduation is defined as one if a child had graduated from school, zero if still attending or dropped out, conditional on enrolled in primary school. I use relatively early educational outcomes as measurements of child quality because the economic theory of the Q-Q trade-off analyses parental choices. It is more convincing to use outcomes that take place closer in time to parental choice-making on quantity and quality to test the theory (Lee, 2008). Therefore, college enrolment and labour market outcomes are not my outcomes of interest; they are more likely to be affected by various factors that are not correlated with parental choices.

[1]Divorce was considered both immoral and bourgeois during 1950s to 1970s, and it required approval from the state, which made it basically impossible. In 1980s, a new marriage law made divorce relatively easier but the divorce rate remained very low, at less than 0.1%.

[2]This is the only information on migration history that the 1990 census provided. China had very rigid hukou registration system that made migration very difficult. Before 1990s, the household mobility in China was almost zero. Indeed, 98% of the households reported to live in the same county and 99% reported to live in the same province during the last 5 years.

[3]This study focuses on public school which was the only type of primary and secondary school that existed in China during the sample period. The number of private school expanded in 2003 due to a change in the education law that admitted the legitimacy of non-government education.

The sample is restricted to children aged between 6 and 16 to ensure that all children have satisfied the minimum entry age for primary school and that there is no adult children living in the house since they are not traceable in the survey. The sample is further restricted to children age smaller than or equal to 13[1] for estimation of the effect on enrolment of primary school, and restricted to children age older than 13 for estimation of completing primary school and junior high school enrolment. Overall, this paper focuses on nine subsamples, which are categorized by family size and outcome variable.

A set of family demographic characteristics are controlled for. In the 1+ sample, the age of the firstborn, mother's age at first birth, education and ethnic group of parents, an rural indicator and provincial indicators are included as controls. It is important to control for maternal age at first birth due to its correlation with twin births and investment behaviour. I also control for cohort effects by including the birth year of the firstborn child and father. It will help address the problem that younger children are more likely to be in school and have few siblings (Qian, 2009). Parental cohort effects are also included in the sense that parents of first-born children are likely to be different from parents of other parity born children. In addition, it controls for the year specific effects. For example, one would expect certain differences between parent who experienced the cultural revolution[2] and those who did not. In the 2+ sample, the gender of firstborn child and birth spacing between first two births are additionally controlled for. In the 3+ sample, I also control for the sex composition of the first and second borns, birth spacing between the second and third born children.

As shown in table 2.3, the primary school enrolment rate is 86% for firstborn in 1+ sample, 86% for firstborn in 2+ sample, 84% for firstborn and 84% for secondborn in 3+ sample. Primary graduation rate is relatively lower, 75% for 1+ sample, 74% for 2+ sample, 74% for firstborns and 68% for secondborns in 3+ sample. Following the similar pattern, the rate of junior school enrolment drops quite a lot in all subsamples. While older siblings have similar enrolment rate of primary school, the rate of primary school graduation and secondary school enrolment decreases with household size. It indicates a negative correlation between family size and child schooling outcome, but not necessarily a causal one. It may simply due to the fact that children of smaller

[1]The years of primary school education varied between 5 and 6 years. Most of the children were expected to complete primary school by the age of 13, conditional on enrolled in primary school and no drop-out.

[2]The Cultural Revolution lasted from 1966 to 1976, and it caused a large scale eduction system disruption.

families are from relatively younger cohorts and benefit from better education system.

The gender of the first birth is biologically normal(0.51) in the 1+ sample, and it decreases in 2+ sample and declines further in the 3+ sample. In addition, sex ratio increases with birth order[1]. The sex ratio at different sample and birth order provides evidence for son preferences. The twinnng rate at each birth is within the natural range[2].

Table 2.4 shows that children are relatively older and mothers are younger at first child birth in households with higher fertility, while mother's and father's age at census are quite similar across subsamples. Parents have higher education level are associated with fewer children. There are smaller proportion of parents of Han ethnicity and from urban areas in higher fertility families.

## 2.5 Empirical Results

This section tests the Q-Q trade-off for the firstborn child in the 1+ sample, 2+ sample and 3+ sample, respectively; for the secondborn child in the 3+ sample. Four specifications are used for the test in each sample, the author will start the estimation with ordinary least square (OLS) estimator, then move to two stage least square (2SLS) estimator. Since 2SLS estimator deals with the endogenous issue using two stages of linear estimation, this paper considers two additional non-linear estimators–nonlinear first-stage(NLIV) and generalized method of moments(GMM). Due to the forbidden regression problem mentioned earlier, NLIV uses Poisson regression to estimate the first stage effects on fertility and then uses it as an instrument for fertility in a linear regression of outcome variable. GMM estimator employs non-linear estimations at both firth and second stage, i.e. obtaining Poisson estimator for fertility and then use it as instrument for fertility in a binary regression for the schooling outcome variable.

### 2.5.1 At Least One Birth Sample

Panel a of table 2.5 shows the effect of a marginal child at second birth on the education outcomes of the firstborn. All the specifications provide evidence on significant Q-Q

[1] Yi et al. (1993) show that the sex ratio of the first child is biologically balanced, but it goes up at higher birth orders especially for households with no previous boy birth. Almond et al. (2013) confirm that the sex ratio of the first child is at natural rate 1.05, but the sex ratio of the second child quite differs, with first girl families have persistently higher rate than families with a first boy.

[2] Huang et al. (2016) show that twins rate is around 0.4% for households with children born between 1974-1980, and slightly higher 0.45% in the period 1980-1984.

trade-off, irrelevant of the outcome variables. OLS estimation suffers from endogeneity problem that parents with different preference for family size may also have different preference for quality of children, and unobserved child's ability may affect parents' investment decision and child's performance. The exogenous variation of sibling size is induced by gender of the first child in the 1+ sample.

In all the cases, 2SLS estimator gives larger Q-Q trade-off than OLS, which implies that the OLS estimator is downward biased. NLIV estimator gives even larger effects than the 2SLS estimator, this is mainly due to the difference in the distribution assumption imposed on fertility. 2SLS assumes that fertility is normal distributed, and NLIV uses Poisson distribution to model fertility. Comparing to the NLIV estimator, GMM estimator gives a slightly greater effect of family size on the probability of enrolling in primary school, and a smaller effect on the probability of graduating from primary school and enrolling in junior school. For example, GMM estimates shows that an additional birth induced by a firstborn girl decreases the probability of primary school enrolment by 19 percent [1], which is 1% higher than the NLIV estimator [2]. The different magnitude of Q-Q trade-off between NLIV and GMM is resulting from the non-linear distribution assumption imposed on the outcome variable. For example, high density of children's outcome are located at the upper tail of the distribution of the nonlinear model, thus it gives better approximation to the data generating process than the linear probability model.

Note that the magnitude of Q-Q trade-off is considerably higher in the case of junior school enrolment. OLS estimator states that the probability of enrolling in junior school for the firstborn is reduced by 11.1%, and 2SLS suggests a reduction in the probability by half. NLIV gives slightly higher estimate of 61.5%, while GMM estimator shows a 54.3% reduction in the likelihood of enrolling in the junior school if there is a marginal second birth.

By exploring the exogenous variation in family size that is induced by first child's gender, this paper is able to fill the gap of examining the effect of family size increasing from 1 to 2 on firstborn. It provides the technical foundation for family planning policies, particularly it may imply that China's One Child Policy is actually beneficial for child's human capital by restricting the family size, thus reducing resource competition. Furthermore, if policy maker aims at encouraging child birth, then corresponding policy

[1] Note that column (4) of panel a gives estimates of -0.162, which corresponds to a 19%(0.162/0.857) reduction in the probability of primary school enrolment for the firstborn in 1+ sample.

[2] Column (3) of panel a gives NLIV estimates of -0.15, which corresponds to a 18% reduction.

(for example, education and health) reform is needed to cooperating with the significant Q-Q trade-off.

### 2.5.2 At Least Two Births Sample

Panel b of table 2.5 gives the estimation results in the households with at least 2 births. All the OLS estimations show significantly negative effects of family size on the outcome of firstborn children, which are likely to be spurious inverse relationships. An additional child at second birth decrease the probability of enrolment by 3%, the probability of primary graduation by 7% and the likelihood of enrolling in junior school by 12.5%. However, OLS estimation neglects unobserved endowment and parental heterogeneity in preferences for quality of children and quantity of children.

By using twins at second birth as an instrumental variable, the conventional 2SLS estimations show smaller effects in magnitude but no evidence of significant Q-Q trade-off, which is in line with previous literature that using the same methodology. The twins instrument generates an average treatment effect on the nontreated, thus the estimator can be interpreted as the impact of an additional child at second birth on the educational attainment of firstborn child in two singleton births family(non-twins family). Note that 2+ sample shows a different direction of OLS bias than the 1+ sample. Comparing to 2SLS, OLS underestimates the true trade-off on the first born in at least 1 birth family while overestimates the effects in 2+ sample. A marginal child at second birth insignificantly increases the probability of primary school enrolment by 0.6%, the probability of primary graduation by 6.1% and of junior enrolment by 4.4%, for the firstborn.

Move to the NLIV estimates, the Q-Q trade-off turns significant, the effects are -2%, -20% and -35%, respectively. GMM estimation supports a significant Q-Q trade-off by accounting for the non-normal distribution of both outcome and endogenous variables. The schooling outcome of firstborn child is reduced by 8.3%, 12.2% and 15.5%, respectively for the probability of primary enrolment, primary graduation and junior enrolment. This finding contrasts the majority of previous literature which finds no Q-Q trade-off, but somehow in line with the results of Rosenzweig and Zhang (2009). They use a logit model and show that twinning reduces the likelihood of college attendance by 27%.

### 2.5.3 At Least Three Births Sample

Estimation results from the at least three births sample give a different picture than the at least two births sample. Panel c of table 2.5 shows effects of a marginal fourth child on firstborn and secondborn children, separately. As it in the case for 2+ sample, OLS estimators show significant negative effects of an extra child on the outcome of elderly siblings, with the effect is slightly greater in magnitude on the secondborn than the firstborn.

This paper examines the effects on the firstborn and secondborn separately, because children of different birth orders experience different degrees of sibling interaction and family resource allocation. Generally the 2SLS estimates show no significant trade-off, with one exception that the probability of enrolling in junior school is significantly reduced by 22.4 percent for the secondborn child. NLIV estimate shows greater effects in magnitude than 2SLS. It gives insignificant Q-Q trade-off on primary school enrolment, and significant trade-off on the other two measurements of schooling outcome. GMM estimator shows that the only significant adverse effect on the firstborn children is the 3.3 percent reduction on the likelihood of primary school graduation, while there is no significant trade-off on enrolment of primary school and junior high school. The secondborn sibling suffers more of an additional fourth child than the firstborn sibling. An additional child decreases the probability of primary school enrolment by 5.5%, primary school graduation by 8.4%, and junior high school enrolment by 10.1% for the secondborn children. This is consistent with Black et al. (2005) in that they find a monotonic decline in educational attainment as the birth order goes up.

It is worth noting that in the summary statistics, firstborn children in 3+ sample have similar rate of primary school enrolment with 1+ and 2+ sample, but lower rate of primary school graduation and junior high school enrolment. It may imply that 3+ families are more financially constrained if resource competition is more intense in families with more children. Families make different decisions on whether to send kids to junior high school or not, it could be the case that more affluent families have higher propensity sending kids to junior high school. Therefore the trade-off is less obvious for those families who are less likely to financially affected by the additional fourth child.

Comparing the trade-off effects in 1+, 2+ and 3+ sample, I find that the trade-off is more evident in a relatively smaller family. It maybe due to the fact that conditional on a bigger family size, the resource is spread more thinly and thus the marginal decrease in the welfare of older sibling declines with child quantity. In other words, for those families who prefer to have a larger family size, an extra child affect the division of family resources less as family size goes up. The smaller magnitude of the estimators obtained for a larger family size provides some evidence on the non-linearity of the trade-off as quantity goes up. In addition, there is evidence of birth order effect. Generally, Q-Q trade-off is greater for the secondborn than for the firstborn in at least 3 births family. Note that the majority of the sample is consists of rural households, thus the estimation is more reflexive for less developed part of China.

### 2.5.4 Robustness Check

#### 2.5.4.1 Instruments Validity

There are concerns about the validity of the instruments if the gender of firstborn and twins affect the schooling outcome independently from family size.

Twinning may affect the outcome of lower births children through zero spacing, lower birthweight and economies of scale. The spacing channel on the outcome of the firstbirth can be tested through 3+ sample without twins but with two tightly spaced following births (Black et al., 2005). Panel (a) of table 2.6 shows that spacing has nearly zero effect on the outcome of older sibling. If this can be extrapolated to the case of twins, then twins will not affect outcome of older sibling independently through family size.

Alternatively, the validity of twins instrument can be tested as proposed by Angrist et al. (2010), they estimate reduced forms in the no-first-stage sample. The idea is that a valid instrument should not affect outcome independently of family size, therefore by shutting down the channel of family size, twin births should have no effect on outcome. The no-first-stage sample consists of families who are likely to have a bigger family anyway, regardless of the twins birth. Indicators for preferring a larger quantity of children could be a tighter spacing of subsequent births and an younger age of mother at first child birth. Panel (b) of table 2.6 shows results based on no-first-stage samples, there is no evidence to reject the validity of twin births as instruments.

Rosenzweig and Zhang (2009) propose that birthweight deficit of twins may induce parents to allocate resources toward lower birth order non-twins children, if this reinforcement story is true, then the estimators obtained through twins at second and third births will be biased positively. My data provides no information on birthweight, although birthweight itself is endogenous. This paper tests the Q-Q trade-off following the bounding strategy proposed by Rosenzweig and Zhang(2009),

$$Y_{ij} = \alpha_1 T_j + \alpha_2 F_{ij} + \alpha_3 (T_j * F_{ij}) + X_i \theta + e_{ij} \tag{2.10}$$

where $T_j$ is an indicator for a household $j$ with twin at second(third) birth, $F_{ij}$ is a dummy for singleton first-born(singleton first two births) and their interaction are included in the regression. The variable of interest is the interaction term, with a positive coefficient $\alpha_3$ implying parental reinforcement behaviour. I observe reinforcement effects toward singleton birth child in the at least two births sample, while the effects are not evident in the at least three births sample (table 2.7).

Following the idea of Rosenzweig and Zhang (2009), the effects of twinning on non-twins older sibling provide the lower bound, and the effects of twinning on twins themselves provide the upper bound. As it indicated by table 2.7, the Q-Q trade-off obtained through twins at second birth gives the lower bound, it would be informative to also know the upper bound by using twins at first birth. Twins at first birth generates bigger effects of fertility on schooling outcome (table 2.8). These estimates are big in magnitude but comparable to the previous finding of Rosenzweig and Zhang (2009) where they find a 20 percent decrease in expected college enrolment and literature grade. Since twining reduces birthweight which is positively correlated with the reinforcement behaviour, the estimations without conditional on birthweight are negatively biased. There is caveat using twin at first birth as an instrumental variable. Twins themselves are not directly comparable to non-twins due to birthweight deficit, the reader should bear in mind that the true Q-Q trade-off effect is possibly overestimated in the case of using twins at first birth as instrumental variable.

Note that estimation based on twins at first birth is not directly comparable to twins at higher birth order and gender of the first birth for three reasons. First, twins at first birth identifies the effect of an additional birth on the average outcome of twins compare to the outcome of singleton birth, while twins at higher birth orders gives the effects of an additional birth on singleton births. Second, twins at first birth is more likely to be an exogenous shock in the sense that parents have no prior on their children, therefore they are less likely to have reinforcement or compensate behaviour. On the contrary, twins at higher birth order may induce parents who value the quality of children to allocate the resource away from twins, which partly offsets the true Q-Q trade-off. Third, different instruments generate different complier population and capture different local average treatment effects(LATE) (Angrist et al., 2010). Twins at first birth affect family size at the incidence of twinning, while girl at first birth affects fertility at a relatively bigger range for families have son preferences and the range will be wider if this preference is stronger. Overall, twins at first birth provide the upper bound of Q-Q trade-off as suggested by Rosenzweig and Zhang (2009), which is consistent with the finding that estimators generated by twins at first birth are bigger in magnitude compares to the case of twins at higher birth order and gender at first birth. As Angrist et al. (2010) point out, different instrument is suffering from different omitted variable bias, therefore comparing the estimation results from two different instrumental strategies would provide a specification check itself.

An additional issue of using twins as instrument is the possibility that twins may affect outcome through economies of scale, depending on the tighter spacing and the gender composition of twins birth. If twins of the same gender benefit from sharing clothes and rooms, then one would expect smaller trade-off to be found in the same gender twins. It twins of the mixed gender benefit from the interaction of different-gender sibling, then a smaller trade-off would be found in the mixed gender twins. The best scenario for twins to be a valid instrument is to have estimation results irrelevant of gender composition of twins. I compare results based on two subsamples of 1+ sample. The same-sex subsample is consists of singleton boy(girl) at first birth and same-sex twins at first birth; the mixed-sex subsample is comprised of singleton at first birth and mixed-sex twins at first birth. If economies of scale do exist, then the formal subsample will be expected to exhibit significantly less Q-Q trade-off. Table 2.9 shows no evidence that twins affect outcome variable directly through economies of scale. There are no significant differences between two subsamples among all outcome variables. For example, the first two columns show the slope of regression line -0.194 and -0.210, clearly their confidence interval are largely overlapped. Therefore, it seems that the internal validity of twins instrument is not threatened in this context.

There is also concern about using the gender of firstborn as instrument for fertility, it may affect outcomes through parental preference in terms of resource allocation. A direct test is to compare the average outcomes of children by number of boys in the family and family size (Lee, 2008). Panel a of table 2.10 shows that households with more boys have higher probability enrolling in primary school conditional on family size, however, the probability of completing primary school and enrolling in secondary school decreases with more boys in the household. This can be regarded as evidence that there is no systemic gender bias in parental investment. Note that this comparison is intuitive, while it does not account for any endogeneity. If parents prefer to invest more in boy rather than girl, then the trade-off is likely to be positively biased for boys and negatively biased for girls. Additionally, panel b of table 2.10 performs a reduced form test in the no-first-stage sample. Indicators for preferring a larger quantity of children could be tighter spacing of subsequent births, younger mother at first child birth and one of the parent belongs to ethnic minorities[1]. Unfortunately, the author fails to rule out the direct impact of first child's gender on education outcomes, but the direction of the bias can be inferred. Since there is a negative correlation between girl birth and outcome variables, the true Q-Q trade-off is overestimated.

#### 2.5.4.2 Heterogeneous Effects

There is some evidence in the literature that the quantity-quality effect differs by families' socioeconomic background. In order to check whether it also applies to this study, I identify potential heterogeneous effects by stratifying the sample by mother's education level. I also test the sensitivity of results to other stratifications of the sample, including rural-urban gap and gender gap. For the ease of presentation, this subsection only reports results based on GMM estimates. In addition, following the logic of bounding strategy (Rosenzweig and Zhang (2009)), GMM estimation based on twins at first birth is also shown in this subsection, despite its problem mentioned above[2].

To identify the heterogeneity in terms of mother's education level, children have been grouped into two sub-samples, depending on whether mothers have completed primary school. The Q-Q trade-off is more evident in families with lower educated mother in 1+ and 2+ samples (table 2.11). In fact, the big negative effects found in table 2.5 are mainly driven by households with low educated mothers. There is relative smaller or no significant negative effects of family size on primary school graduation and junior high school enrolment for children of high educated mothers. However, estimation results turn opposite in the case of at least three births family. The Q-Q trade-off is not subject to mothers' education level when the outcome is measured by primary school enrolment, but children in high-educated families turn out to suffer

[1] Ethnic minorities are waved from the One Child Policy. I would expect girl at first birth has less impact on the family size in the non-han sample than the Han-sample.

[2] There is caveat using twin at first birth as an instrumental variable. On one hand, twins themselves are not directly comparable to non-twins due to birthweight deficit, therefore, twins at first birth provides the upper bound of Q-Q trade-off. On the other hand, twins at first birth is more likely to be an exogenous shock to the family. Parents have no prior on their children, therefore they are less likely to have reinforcement or compensate behaviour. On the contrary, twins at higher birth order may induce parents who value the quality of children to allocate the resource away from twins, which partly offsets the true Q-Q trade-off.

more of the adverse fertility effects than children in low-educated families in the case of primary graduation and junior high school enrolment. Possible explanation could be drawn from mother's labour force participation. I find that mother's employment status keeps stable when family size increase from 2 to 3, but decreases significantly when it increases from 3 to 4. If education level could proxy the wealth of the family, then children of less educated mothers suffer from more severe resource competition, that explains why the Q-Q trade-off is more evident for children of less educated mothers in the 1+ and 2+ sample. However, since mothers reduce labour supply in 3+ sample, then education level is a poor indicator for wealth effect in this case. Conditional on enrolled in primary school, children of high educated mothers have much higher probability of graduation rate and enrolment rate of junior school[1]. However, the adverse effect of a marginal fourth child is bigger conditional on a high-educated family than conditional on a low-educated family.

Rural households are more likely to be financially constrained and the education system in rural areas are relatively poorer. Therefore, it is reasonable to consider that the fertility effects would vary by rural and urban areas. Table 2.12 presents considerable adverse effects on children's outcome in rural areas. The significant Q-Q trade-off found in the main regression largely attributes to children of rural households, especially rural households account for a large fraction of my sample. The rural-urban gap shows a different pattern for first born and second born children in families with at least three births. There is no significant difference found for the first born children, but a rural-urban gap is observed in the probability of graduating from primary school for the second born children. A marginal fourth child significantly decreases the probability of completing primary school for the second born by 5.1 percentage points in rural areas, while there is no trade-off found for the urban areas.

If families have strong son preferences, then an additional girl birth will affect the performance of older siblings to a less extent than an extra boy birth. I constructed two sub-samples for 2+ sample—one sub-sample excludes households with twin girls at second birth to ensure that one of the twins must be a boy, and the other sub-sample excludes households with twin boys at second birth to ensure that one of the twins must be a girl. The same restriction applies to 3+ sample and the subjects of interest are the average outcomes of first two births. Table 2.13 shows no significant gender gap by

[1]Conditional on enrolled in primary school, the probability of graduation is 0.69 for low-educated families and 0.84 for high-educated families; the likelihood of enrolling in junior school is 0.48 for low-educated families and 0.76 for high educated families.

the marginal child, and the magnitude of the effect of an extra boy birth is very similar to an extra girl birth. It could be explained by the possibility that families decide to have larger families have less son preference for gender composition of children. In other words, there is no post-natal gender preference conditional on child birth, parents treat the new born child indifferently of gender. Almond et al. (2013) argue that sex mix is most preferred, conditional on having two children in the family.

I further test Q-Q trade-off by gender composition of older siblings. Table 2.14 shows that firstborn girls suffer more from an additional child birth comparing to firstborn boys in primary school education. While there is no significant differences observed for secondary school enrolment in 2+ sample, results for 3+ sample show that the rate of junior high school enrolment is reduced by 10 percentage points for boys but no impacts on girls. The adverse effect on the primary school education of girls may imply that families with son preferences suppress the education opportunity for girls at an early stage. However, given no significant differences on primary school graduation rate, the enrolment of junior high school may be related with the higher education cost incurred, if boys require higher costs than girls and families are financially constrained, then an additional fourth child affects more negatively on boys.

#### 2.5.4.3 One Child Policy

The overlap of my sample period and the One Child Policy(OCP) raises concern for the credibility of the estimation. McElroy and Yang (2000) finds that the removal of the One Child Policy would increase family size by one third of a child, and Qian (2009) shows that relaxation of the policy in 1982 for rural families with first born girls lead to an increase in fertility by one fourth of a child. Rosenzweig and Zhang (2009) show that the policy would lead to a less than 4 percent increase in educational attainment, a less than 9 percent increase in the expected college enrolment. Liu (2014) suggests the men's income would be increased by 1.71% and women's income by 1.55% at most by One-Child Policy. Overall, these analysis show that the consequences of the One Child Policy on fertility is modest and on human capital accumulation is negligible.

Although sex-selective abortion is not widely available until late 1980s, it is reasonable to suspect validity of instrumental variables as the ultrasound services to some extent allows parents to decide the gender composition of the children. If this bias exist, I would expect the contamination of estimation is less severe in a pre One Child Policy period. My sample period covers the implementation of the One Child Policy since 1979, a sensitivity test is carried out using 1982 census. If parents are allowed to choose the quantity of children, the willingness to trade quality for quantity would indicate a bigger Q-Q trade-off. However, table 2.15 shows that the Q-Q trade-off is smaller in the absence of One Child Policy. There could be two candidate explanations. First, the children in 1982 sample were born between 1966 and 1976, which coincided with the cultural revolution period. This revolution mainly disrupted senior high school and higher education, and primary school and junior high school education system were disorganised. Meng and Gregory (2007) describe that primary school students were not taught in standard curricula and they spent most of their school time doing manual work in factories and the countryside. Although surprisingly the primary school graduation rates(around 90%) were higher than the 1990 sample, it is likely that that there was a change in the graduation criterion, which made the trade-off less evident. Second, it could be that those who violated the OCP for an extra birth suffered from bad economics and socioeconomic situation so that it adversely affected the outcome of children.

## 2.6 Conclusion

This paper studies the effect of fertility on children's human capital accumulation in the context of China. The main contribution comes from applying a novel GMM approach for testing the existence of Q-Q trade-off. This method is more favourable than conventional two stage least squares because it takes into account the limited dependent variable nature of the data. Thus, this paper not only considers the non-linear fertility effect across parities by examining different birth sample, but also accounts for non-linear distribution within parities. While the conventional approach shows no effect of fertility on education performance of children, GMM estimators give significant Q-Q trade-offs. This finding is consistent with the Becker-Lewis model that an increase in child quantity increases the shadow price of quality. Moreover, the trade-off nonlinearly decreases with family size and shows heterogeneous effects by birth order.

This paper is the first to apply first child's gender as an instrumental variable for testing Q-Q trade-off in the context of China where the son preferences are prevailing. This instrument offers the possibility to test the Q-Q trade-off of a marginal second birth on the first born while typically twins and parental preference for a mixed gender composition estimate the effect of a marginal third or higher birth order child on older siblings.

Although arguably endogeneous, this analysis also employs twins at first birth as an instrumental variable for fertility to provide an upper bound estimation for the Q-Q trade-off. Rosenzweig and Zhang (2009) suggest that the upper bound estimator is negatively biased and the lower bound estimator is positively biased, if not accounting for the birthweight deficit of twins and parental reinforcement behaviour. There is no birthweight information available in the census, but I find evidence of parental reinforcement behaviour in the 2+ sample, by comparing the effect of twinning on twins and non-twins. Therefore, estimator in the 2+ sample is regarded as a positively biased lower bound. In other words, the true Q-Q trade-off are bigger in magnitude for the 2+ sample. The upper bound is provided by using twins at first birth. Additionally, the effect of fertility on the outcome of older siblings are provided for first and second birth separately in the 3+ sample, as I believe that the trade-off would differ by birth order due to differences in sibling interaction and parental resource allocation. Indeed, results show heterogeneous birth order effects, which are obscured by previous literature.

Several validity tests are conducted to ensure that the estimates are not driven by direct effects of instrumental variables on outcomes, if any, at least the direction of bias is obtained. Twinning may affect the outcome of lower births children through zero spacing(economies of scale) and lower birthweight. I test the existence of these mechanisms and results suggest that some estimators are probably positively biased. Thus the true adverse effects of family size on the schooling performance of children are bigger in magnitude, with the exception of the effect on twin themselves. Son preferences may affect outcome independently through gender-based parental investment, for instance, families with son preferences will tend to allocate resource toward boys and the Q-Q trade-off will be positively biased for boys and negatively biased for girls. There is evidence that gender of firstborn directly affects outcomes, thus Q-Q trade-off is negatively biased. In other words, the true effects of an additional birth induced by a girl birth on quality is weaker than what the estimators imply.

In addition, I explore heterogeneous effects across family socioeconomics background, rural urban areas, gender of older siblings and extra child. The Q-Q trade-off are mainly driven by children of households from rural areas and low-educated parents in the sample of at least one or two births, while the trade-off is more obvious in high-educated parents in a larger family size. Results based on gender confirms the finding of Almond et al. (2013) that parents prefer sex mix than 2 boys, followed by 2 girls.

This study provides technique support for policies aiming at reducing contraceptive costs, controlling population growth and subsidizing families with fewer children. The estimates show that those policies may in fact contribute to better educational outcomes of children. However, population control is unsustainable for long term economic growth especially in ageing societies. If policy maker aims at alleviating population ageing problem by encouraging child birth, then corresponding policy reform is needed to cooperate with the adverse fertility effect. For example, maternal leave, childcare provision, public education and health system should be adjusted accordingly.

# Tables

Table 2.1: Relevant Test: First Child's Gender

| | Parity Progression | | | Weibull Hazard Model of Childbirth Spacing | | |
|---|---|---|---|---|---|---|
| | No.children>1 | No.children>2 | No.children>2 | spacing12 | spacing23 | spacing23 |
| girl1 | 0.055*** | 0.095*** | 0.005 | 1.564*** | 1.539*** | 0.946*** |
| | (0.003) | (0.003) | (0.003) | (0.007) | (0.014) | (0.013) |
| girl2 | | 0.129*** | | | 1.625*** | |
| | | (0.003) | | | (0.015) | |
| samesex12 | | 0.112*** | | | 1.197*** | |
| | | (0.003) | | | (0.011) | |
| boy12 | | | -0.001 | | | 0.736*** |
| | | | (0.003) | | | (0.011) |
| girl12 | | | 0.094*** | | | 1.948*** |
| | | | (0.003) | | | (0.022) |
| N | 60437 | 60437 | 58883 | 566133 | 222428 | 222428 |

Note: All the number contained in the variable name indicates birth order, for instance, spacing12 means spacing between first and second births. Robust standard errors are reported in parentheses.

Table 2.2: First Stage Effects on Fertility

**Panel a: 1+ sample**

| Sub-sample | Primary enrolment | | Primary graduation | |
|---|---|---|---|---|
| **Instrument** | **twin1** | **girl1** | **twin1** | **girl1** |
| No. Children | 0.513*** | 0.275*** | 0.372*** | 0.298*** |
| | (0.009) | (0.002) | (0.025) | (0.004) |
| Wald test | 3280.220 | 23884.040 | 221.960 | 5140.730 |
| N | 610484 | 606422 | 148969 | 148303 |

**Panel b: 2+ sample**

| Sub-sample | Primary enrolment | Primary graduation |
|---|---|---|
| **Instrument** | **twin2** | **twin2** |
| No. Children | 0.561*** | 0.562*** |
| | (0.008) | (0.027) |
| Wald test | 4899.800 | 425.430 |
| N | 429120 | 137560 |

**Panel c: 3+ sample**

| Sub-sample | Primary enrolment | Primary graduation |
|---|---|---|
| **Instrument** | **twin3** | **twin3** |
| | | |
| **First born** | | |
| No. Children | 0.426*** | 0.523*** |
| | (0.011) | (0.027) |
| Wald test | 1441.900 | 389.080 |
| N | 153693 | 69735 |
| | | |
| **Second born** | | |
| No. Children | 0.482*** | 0.547*** |
| | (0.013) | (0.032) |
| Wald test | 1411.370 | 295.600 |
| N | 163451 | 58141 |

Note: Panel a shows the first-stage effects of twins and girl at first birth on fertility for the 1+ sample. Panel b shows the first-stage effects of twins at second birth on fertility for the 2+ sample. Panel c shows the first-stage effects of twins at third birth on fertility for the 3+ sample.

All the number contained in the variable name indicates birth order, for instance, girl1 indicates girl at first birth and twin2 means twins at second birth.

Regressions for panel a include controls for child's age, gender, mother's age at first child birth, father's age and year of birth, panel b also controls for mother's age at census, birth spacing between first and second birth, panel c additionally controls for sex composition of the first two birth, birth spacing between the second and third birth. Robust standard errors are reported in parentheses and are clustered at household level for the 3+ sample.

Table 2.3: Descriptive Statistics by Sample 1/2

| | 1+ sample | | 2+ sample | | 3+ sample | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | First born | | Second born | |
| | mean | std dev | mean | std dev | mean | std dev | mean | std dev |
| **Endogeneous variable** | | | | | | | | |
| No. children | 2.152 | 0.935 | 2.531 | 0.764 | 3.336 | 0.629 | 3.416 | 0.703 |
| **Education outcomes** | | | | | | | | |
| Primary Enrolment ($6 <= age <= 13$) | 0.857 | 0.351 | 0.865 | 0.342 | 0.841 | 0.366 | 0.841 | 0.366 |
| **N** | 610484 | | 429120 | | 153693 | | 163451 | |
| Primary Completed ($13 < age <= 16$) | 0.752 | 0.432 | 0.743 | 0.437 | 0.681 | 0.466 | 0.653 | 0.476 |
| Junior Enrolment ($13 < age <= 16$) | 0.602 | 0.490 | 0.586 | 0.493 | 0.485 | 0.500 | 0.456 | 0.498 |
| **N** | 148969 | | 137560 | | 69735 | | 58141 | |
| **Family Composition** | | | | | | | | |
| twin1 | 0.006 | 0.079 | | | | | | |
| twin2 | | | 0.004 | 0.063 | | | | |
| twin3 | | | | | 0.003 | 0.053 | 0.003 | 0.052 |
| boy1 | 0.512 | 0.499 | 0.482 | 0.500 | 0.409 | 0.492 | 0.422 | 0.494 |
| boy2 | | | 0.526 | 0.499 | 0.420 | 0.494 | 0.433 | 0.496 |
| boy3 | | | | | 0.541 | 0.498 | 0.531 | 0.499 |
| boy12 | | | 0.241 | 0.428 | 0.198 | 0.399 | 0.206 | 0.404 |
| girl12 | | | 0.234 | 0.423 | 0.369 | 0.482 | 0.351 | 0.477 |
| boy123 | | | | | 0.087 | 0.282 | 0.092 | 0.289 |
| girl123 | | | | | 0.141 | 0.348 | 0.141 | 0.348 |
| N | 759453 | | 566680 | | 223428 | | 221592 | |

Note: 1+ sample includes households with at least 1 child birth(note here 1+ sample is the sample I use for estimation with twinning at first birth as instrument. When first child' sex is used as instrumental variable, the 1+ sample excludes families with twins at first birth, and all the summary statistics remains largely unchanged); 2+ sample is consists of families with at least 2 child births, 3+ sample is comprised of households with at least 3 births.
For the ease of table presentation, it reports the summary statistics for aggregated sample of primary school enrolment and junior school enrolment. However, in the regression analysis, sample differs by children exposed to the outcome. Specifically, in the case of primary enrolment, the sample only contains children age between 6 and 13; in other cases, sample is comprised of children age above 13 but not older than 16.
All the number contained in the variable name indicates birth order, for instance, twin1 means twins at first birth, spacing12 means spacing between first and second births.

Table 2.4: Descriptive Statistics by Sample 2/2

| | 1+ sample | | 2+ sample | | 3+ sample | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | First born | | Second born | |
| | mean | std dev | mean | std dev | mean | std dev | mean | std dev |
| **Control Variables** | | | | | | | | |
| child's age | 10.194 | 3.110 | 10.661 | 3.168 | 11.240 | 3.229 | 11.022 | 3.037 |
| child's birth year | 1979.260 | 3.143 | 1978.784 | 3.198 | 1978.198 | 3.261 | 1978.444 | 3.076 |
| mother's age at first child birth | 23.814 | 3.025 | 23.334 | 2.844 | 22.920 | 2.835 | 22.697 | 2.792 |
| mother's age at census | 34.008 | 4.143 | 33.995 | 4.147 | 34.160 | 4.090 | 36.180 | 3.734 |
| father's age at census | 36.153 | 4.663 | 36.196 | 4.647 | 36.399 | 4.645 | 38.489 | 4.462 |
| *Mother Education level* | | | | | | | | |
| illiterate | 0.359 | 0.480 | 0.426 | 0.494 | 0.510 | 0.500 | 0.544 | 0.498 |
| primary school completed | 0.311 | 0.463 | 0.346 | 0.476 | 0.333 | 0.471 | 0.335 | 0.472 |
| junior high school completed | 0.221 | 0.415 | 0.173 | 0.378 | 0.127 | 0.333 | 0.100 | 0.301 |
| senior high completed | 0.099 | 0.298 | 0.054 | 0.226 | 0.030 | 0.171 | 0.020 | 0.141 |
| college completed | 0.009 | 0.095 | 0.002 | 0.048 | 0.000 | 0.016 | 0.000 | 0.021 |
| *Father Education level* | | | | | | | | |
| illiterate | 0.158 | 0.364 | 0.183 | 0.386 | 0.216 | 0.412 | 0.236 | 0.425 |
| primary school completed | 0.320 | 0.466 | 0.359 | 0.480 | 0.368 | 0.482 | 0.403 | 0.491 |
| junior high school completed | 0.344 | 0.475 | 0.322 | 0.467 | 0.302 | 0.459 | 0.271 | 0.444 |
| senior high school completed | 0.153 | 0.360 | 0.127 | 0.333 | 0.112 | 0.315 | 0.086 | 0.280 |
| college completed | 0.025 | 0.157 | 0.009 | 0.095 | 0.002 | 0.046 | 0.004 | 0.060 |
| mother ethnicity(Han) | 0.931 | 0.254 | 0.917 | 0.276 | 0.884 | 0.320 | 0.889 | 0.314 |
| father ethnicity(Han) | 0.930 | 0.254 | 0.917 | 0.275 | 0.887 | 0.317 | 0.891 | 0.312 |
| mother's birth year | 1955.477 | 4.139 | 1955.491 | 4.142 | 1955.332 | 4.082 | 1953.309 | 3.725 |
| father's birth year | 1953.323 | 4.660 | 1953.281 | 4.643 | 1953.084 | 4.640 | 1950.991 | 4.457 |
| spacing12 | 37.625 | 22.703 | 37.808 | 22.580 | 30.525 | 15.104 | 29.708 | 13.583 |
| spacing23 | | | | | 35.338 | 20.610 | 38.324 | 22.938 |
| rural | 0.775 | 0.418 | 0.894 | 0.308 | 0.955 | 0.208 | 0.938 | 0.242 |
| N | 759453 | | 566680 | | 223428 | | 221592 | |

Note: Same note as it stated in table 2.3.

Table 2.5: The Effect of Family Size on Education Outcome

| | (1) OLS | (2) 2SLS[a] | (3) NLIV[b] | (4) GMM[c] | (5) OLS | (6) 2SLS | (7) NLIV | (8) GMM |
|---|---|---|---|---|---|---|---|---|
| | **Panel a: 1+ sample** | | | | **Panel b: 2+ sample** | | | |
| | **IV: girl at first birth** | | | | **IV: twin at the second birth** | | | |
| Primary Enrolment | | | | | | | | |
| No. kids | -0.019*** | -0.078*** | -0.150*** | -0.162*** | -0.023*** | 0.005 | -0.015*** | -0.072*** |
| | **-0.022**[d] | **-0.091** | **-0.175** | **-0.189** | **-0.027** | **0.006** | **-0.017** | **-0.083** |
| | (0.001) | (0.003) | (0.003) | (0.002) | (0.001) | (0.008) | (0.004) | (0.004) |
| N | 606422 | 606422 | 606422 | 606422 | 429120 | 429120 | 429120 | 429120 |
| Primary Graduation | | | | | | | | |
| No. kids | -0.041*** | -0.089*** | -0.163*** | -0.115*** | -0.049*** | 0.045 | -0.151*** | -0.091*** |
| | **-0.055** | **-0.118** | **-0.217** | **-0.153** | **-0.066** | **0.061** | **-0.203** | **-0.122** |
| | (0.001) | (0.007) | (0.007) | (0.009) | (0.001) | (0.030) | (0.009) | (0.011) |
| N | 148303 | 148303 | 148303 | 148303 | 137560 | 137560 | 137560 | 137560 |
| Junior Enrolment | | | | | | | | |
| No. kids | -0.067*** | -0.300*** | -0.370*** | -0.327*** | -0.073*** | 0.026 | -0.203*** | -0.091*** |
| | **-0.111** | **-0.498** | **-0.615** | **-0.543** | **-0.125** | **0.044** | **-0.346** | **-0.155** |
| | (0.001) | (0.008) | (0.008) | (0.005) | (0.002) | (0.034) | (0.009) | (0.014) |
| N | 148303 | 148303 | 148303 | 148125 | 137560 | 137560 | 137560 | 137394 |
| | **Panel c: 3+ sample** | | | | | | | |
| | **First Born** | | | | **Second Born** | | | |
| | **IV: twin at third birth** | | | | **IV: twin at third birth** | | | |
| Primary Enrolment | | | | | | | | |
| No. kids | -0.025*** | -0.001 | 0.013 | -0.011 | -0.033*** | -0.008 | -0.001 | -0.046*** |
| | **-0.030** | **-0.001** | **0.015** | **-0.059** | **-0.039** | **-0.010** | **-0.001** | **-0.055** |
| | (0.001) | (0.017) | (0.010) | (0.013) | (0.001) | (0.018) | (0.009) | (0.007) |
| N | 153693 | 153693 | 153693 | 153693 | 163451 | 163451 | 163451 | 163424 |
| Primary Graduation | | | | | | | | |
| No. kids | -0.048*** | 0.044 | -0.057*** | -0.033* | -0.053*** | -0.022 | -0.074*** | -0.055** |
| | **-0.070** | **0.065** | **-0.084** | **-0.048** | **-0.081** | **-0.034** | **-0.113** | **-0.084** |
| | (0.002) | (0.035) | (0.016) | (0.017) | (0.003) | (0.041) | (0.015) | (0.022) |
| N | 69735 | 69735 | 69735 | 69710 | 58141 | 58141 | 58141 | 58141 |
| Junior Enrolment | | | | | | | | |
| No. kids | -0.066*** | 0.036 | -0.045*** | -0.009 | -0.068*** | -0.102** | -0.053*** | -0.046** |
| | **-0.136** | **0.074** | **-0.093** | **-0.019** | **-0.149** | **-0.224** | **-0.116** | **-0.101** |
| | (0.002) | (0.040) | (0.007) | (0.019) | (0.002) | (0.044) | (0.014) | (0.021) |
| N | 69735 | 69735 | 153693 | 69556 | 58141 | 58141 | 58141 | 58054 |

Note: Panel a shows the effects of fertility on schooling outcome of firstborns in the 1+ sample. Panel b shows the effects of fertility on schooling outcome of firstborns in the 2+ sample, panel c shows the effect on schooling outcome of firstborns and secondborns, respectively, in the 3+ sample.

[a] 2SLS gives the two stage least square estimates. [b] NLIV gives the estimates based on non-linear first-stage. [c]GMM gives the result taking into account non-linear distribution in both first and second stage.

[d] Numbers in bold gives the corresponding transformation of the magnitude of Q-Q trade-off, i.e. **-0.022** means that an extra kid decreases the probability of primary enrolment by 2.2%, which corresponds to 1.9 percentage points decrease in the outcome variable compared to its mean value(0.019/0.857). The same calculation applies to all the estimates.

All the regressions include control variables described in the note of table 2.2. Robust standard errors are reported in parentheses.

Table 2.6: Testing the Internal Validity of Instruments — Birth Spacing

**Panel a: spacing channel—Black et al.(2005)**

| | Primary enrolment | Primary graduation | Junior enrolment |
|---|---|---|---|
| spacing23 | -0.000*** | 0.000*** | 0.000** |
| | (0.000) | (0.000) | (0.000) |
| N | 207734 | 207752 | 207218 |

**Panel b: no-first-stage—Angrist et al. (2010)**

| | Primary enrolment | Primary graduation | Junior enrolment |
|---|---|---|---|
| | | **Tight birth spacing sample** | |
| twins at second birth | 0.017 | 0.066 | 0.019 |
| | (0.014) | (0.054) | (0.054) |
| N | 142381 | 40431 | 40394 |
| twins at third birth | -0.018 | -0.149 | -0.177 |
| | (0.027) | (0.091) | (0.119) |
| N | 27846 | 6278 | 6278 |
| | | **Young mother sample** | |
| twins at second birth | -0.028 | 0.010 | -0.037 |
| | (0.019) | (0.044) | (0.051) |
| N | 104272 | 46338 | 46281 |
| twins at third birth | 0.064 | -0.015 | -0.144* |
| | (0.066) | (0.069) | (0.082) |
| N | 26234 | 13530 | 13532 |

Note: In panel a, sample is restricted to non-twin families with two tightly spaced second and third births.
Panel b performs reduced form estimations in no-first-stage samples as proposed by Angrist et al. (2010). Those samples are families with tightly spaced births, or households with young mothers. Robust standard errors are reported in parentheses and are clustered by household ID for 3+ sample.

Table 2.7: Testing the Internal Validity of Instruments — Bounding Strategy

| | **Panel a: 2+ sample** | | |
|---|---|---|---|
| | Primary enrolment | Primary graduation | Junior enrolment |
| twins at second birth | -0.023*** | -0.010 | -0.047** |
| | (0.006) | (0.017) | (0.020) |
| first-birth (non-twin) | 0.004*** | 0.045*** | 0.064*** |
| | (0.001) | (0.003) | (0.003) |
| twins at second birth | 0.024*** | 0.037 | 0.063** |
| * first-birth (non-twin) | (0.009) | (0.029) | (0.032) |
| N | 802109 | 258258 | 258258 |

| | **Panel b: 3+ sample** | | |
|---|---|---|---|
| | Primary enrolment | Primary graduation | Junior enrolment |
| twins at third birth | -0.025 | -0.073 | -0.053 |
| | (0.019) | (0.093) | (0.120) |
| first two births (non-twin) | 0.029*** | 0.067*** | 0.059*** |
| | (0.003) | (0.009) | (0.011) |
| twins at third birth | -0.004 | 0.053 | -0.011 |
| * first two births (non-twin) | (0.028) | (0.093) | (0.120) |
| N | 155408 | 48768 | 48705 |

Note: Panel a applies the bounding strategy (Rosenzweig and Zhang, 2009) to 2+ sample, while panel b performs similar methodology on 3+ sample. Robust standard errors are reported in parentheses and are clustered by household ID for 3+ sample.

Table 2.8: The lower bound of using twin births as instrument

| | (1) OLS | (2) 2SLS[a] | (3) NLIV[b] | (4) GMM[c] |
|---|---|---|---|---|
| **IV: twin at the first birth** | | | | |
| No. childnum | -0.017*** | -0.018* | -0.226*** | -0.203*** |
| | **-0.020**[d] | **-0.021** | **-0.264** | **-0.237** |
| | (0.001) | (0.009) | (0.005) | (0.007) |
| N | 610484 | 610484 | 610484 | 610484 |
| No. childnum | -0.039*** | -0.185*** | -0.380*** | -0.262*** |
| | **-0.052** | **-0.246** | **0.505** | **-0.348** |
| | (0.001) | (0.044) | (0.017) | (0.007) |
| N | 148969 | 148969 | 148969 | 148969 |
| No. childnum | -0.059*** | -0.212*** | -0.563*** | -0.343*** |
| | **-0.098** | **-0.352** | **0.935** | **-0.570** |
| | (0.001) | (0.045) | (0.022) | (0.031) |
| N | 148969 | 148969 | 148969 | 148791 |

Note: Panel a shows the effects of fertility on schooling outcome of firstborns in the 1+ sample. Panel b shows the effects of fertility on schooling outcome of firstborns in the 2+ sample, panel c shows the effect on schooling outcome of firstborns and secondborns, respectively, in the 3+ sample.
[a] 2SLS gives the two stage least square estimates. [b] NLIV gives the estimates based on non-linear first-stage. [c]GMM gives the result taking into account non-linear distribution in both first and second stage.
[d] Numbers in bold gives the corresponding transformation of the magnitude of Q-Q trade-off, i.e. **-0.020** means that an extra kid decreases the probability of primary enrolment by 2%, which corresponds to 1.7 percentage points decrease in the outcome variable compared to its mean value(0.017/0.857). The same calculation applies to all the estimates.
All the regressions include control variables described in the note of table 2.2. Robust standard errors are reported in parentheses.

Table 2.9: Testing the Internal Validity of Instruments — Economies of Scale

| | Economies of scale | | | | | |
|---|---|---|---|---|---|---|
| | Primary Enrolement | | Primary Graduation | | Junior Enrolment | |
| | same-sex twin | mixed-sex twin | same-sex twin | mixed-sex twin | same-sex twin | mixed-sex twin |
| No. children | -0.194*** | -0.210*** | -0.271*** | -0.210*** | -0.402*** | -0.393*** |
| | (0.003) | (0.012) | (0.002) | (0.005) | (0.004) | (0.004) |
| N | 609332 | 607574 | 148765 | 148507 | 148587 | 148329 |

Note: Two sub-samples are used to identify the economies of scale channel. The same-sex twin sample consists of singleton boy(girl) and boy-boy(girl-girl) twinning at first birth. The mixed-sex twin sample is comprised of singleton boy(girl) and boy-girl twinning at first birth. Robust standard errors are reported in parentheses.

Table 2.10: Testing the Internal Validity of Instruments — Son Preferences

**Panel a: education outcome by number of boys and family size**

| | No. boys | | | |
|---|---|---|---|---|
| | Primary enrolment | | | |
| **No. births** | 0 | 1 | 2 | 3 |
| 1 | 0.822 | 0.847 | . | . |
| 2 | 0.878 | 0.885 | 0.887 | . |
| 3 | 0.838 | 0.853 | 0.858 | 0.865 |
| | primary graduation | | | |
| 1 | 0.879 | 0.858 | . | . |
| 2 | 0.841 | 0.807 | 0.791 | . |
| 3 | 0.767 | 0.732 | 0.719 | 0.715 |
| | Junior enrolment | | | |
| 1 | 0.814 | 0.792 | . | . |
| 2 | 0.728 | 0.683 | 0.679 | . |
| 3 | 0.587 | 0.539 | 0.538 | 0.548 |

**Panel b: no-first-stage**

| | Primary enrol | Primary grad | Junior enrol |
|---|---|---|---|
| | **Tight spacing** | | |
| girl at first birth | -0.027*** | -0.018*** | -0.085*** |
| | (0.001) | (0.004) | (0.004) |
| N | 142381 | 40431 | 40394 |
| | **Young mother** | | |
| girl at first birth | -0.041*** | -0.047*** | -0.118*** |
| | (0.002) | (0.004) | (0.004) |
| N | 121039 | 47722 | 47662 |
| | **Ethnic minorities** | | |
| girl at first birth | -0.059*** | -0.055*** | -0.082*** |
| | (0.003) | (0.008) | (0.008) |
| N | 41656 | 10966 | 10809 |

Note: In panel a, average education outcomes of children are listed by number of boys and family size.
In panel b, sample is restricted to families with tightly spaced births, or households with young mothers, or families with Non-Han ethnicity(exempted from the One Child Policy). Robust standard errors are reported in parentheses.

Table 2.11: Robustness Check: Heterogeneity—Q-Q Trade-off by Mother's Education

| | Primary Enrolement | | Primary graduation | | Junior Enrolment | |
|---|---|---|---|---|---|---|
| | low-edu | high-edu | low-edu | high-edu | low-edu | high-edu |
| **Panel a: 1+ sample** | | | | | | |
| **IV: girl at first birth** | | | | | | |
| No. children | -0.124*** | -0.060*** | -0.111*** | 0.032*** | -0.343*** | -0.018 |
| | (0.004) | (0.007) | (0.006) | (0.012) | (0.009) | (0.013) |
| N | 321624 | 107496 | 115425 | 22135 | 115259 | 22135 |
| **IV: twins at first birth** | | | | | | |
| No. children | -0.065*** | -0.052*** | -0.087 | -0.012 | -0.096*** | -0.037 |
| | (0.005) | (0.008) | (0.067) | (0.012) | (0.034) | (0.031) |
| N | 323856 | 109326 | 115867 | 22359 | 115701 | 22359 |
| **Panel b: 2+ sample** | | | | | | |
| **IV: twins at second birth** | | | | | | |
| No. children | -0.042*** | -0.044*** | -0.065*** | -0.003 | -0.069*** | -0.019 |
| | (0.006) | (0.008) | (0.014) | (0.013) | (0.016) | (0.016) |
| N | 321624 | 107496 | 115425 | 22135 | 115259 | 22135 |
| **Panel c: 3+ sample** | | | | | | |
| **IV: twins at third birth** | | | | | | |
| No. children | -0.040 | 0.000 | -0.002 | -0.089* | 0.003 | -0.116** |
| | (0.027) | (0.025) | (0.067) | (0.048) | (0.026) | (0.059) |
| N | 175914 | 32424 | 175914 | 3454 | 175378 | 3454 |

Note: High educated mother is defined as completed primary school. Robust standard errors are reported in parentheses and are clustered by household ID for 3+ sample.

Table 2.12: Robustness Check: Heterogeneity—Q-Q Trade-off in terms of Rural-Urban Gap

| | Primary Enrolement | | Primary graduation | | Junior Enrolment | |
|---|---|---|---|---|---|---|
| | rural | urban | rural | urban | rural | urban |
| | | | **Panel a: 1+ sample** | | | |
| **IV: girl at first birth** | | | | | | |
| No. children | -0.131*** | -0.096*** | -0.136*** | 0.027* | -0.402*** | 0.018 |
| | (0.002) | (0.009) | (0.009) | (0.015) | (0.006) | (0.015) |
| N | 466549 | 139873 | 118935 | 29368 | 118935 | 29368 |
| **IV: twins at first birth** | | | | | | |
| No. children | -0.210*** | -0.084*** | -0.301*** | -0.065*** | -0.339*** | -0.092*** |
| | (0.004) | (0.009) | (0.003) | (0.009) | (0.010) | (0.011) |
| N | 469207 | 141277 | 119345 | 29624 | 119345 | 29624 |
| | | | **Panel b: 2+ sample** | | | |
| **IV: twins at second birth** | | | | | | |
| No. children | -0.082*** | -0.018* | -0.110*** | -0.021*** | -0.098*** | -0.048*** |
| | (0.005) | (0.011) | (0.014) | (0.008) | (0.017) | (0.014) |
| N | 392259 | 36861 | 114267 | 23293 | 114267 | 23293 |
| | | | **Panel c: 3+ sample** | | | |
| **IV: twins at third birth − average outcome of first two births** | | | | | | |
| No. children | -0.054*** | -0.061*** | -0.015 | -0.073 | -0.053 | -0.029 |
| | (0.012) | (0.021) | (0.030) | (0.066) | (0.037) | (0.069) |
| N | 200862 | 7476 | 30166 | 3584 | 30166 | 3584 |
| **IV: twins at third birth** | | | | | | |
| **1st birth** | | | | | | |
| No. children | -0.081*** | -0.064** | -0.008 | -0.032 | 0.020 | 0.002 |
| | (0.011) | (0.032) | (0.021) | (0.031) | (0.022) | (0.040) |
| N | 148754 | 4939 | 64529 | 5206 | 64529 | 5206 |
| **2nd birth** | | | | | | |
| No. children | -0.079*** | -0.048*** | -0.051** | -0.015 | -0.039 | -0.017 |
| | (0.010) | (0.017) | (0.022) | (0.028) | (0.025) | (0.033) |
| N | 155822 | 7629 | 51932 | 6209 | 51932 | 6209 |

Note: Robust standard errors are reported in parentheses and are clustered by household ID for 3+ sample.

Table 2.13: Robustness Check: Heterogeneity—Q-Q Trade-off by Gender of the Extra Sibling

| | Primary Enrolement | | Primary graduation | | Junior Enrolment | |
|---|---|---|---|---|---|---|
| | boy | girl | boy | girl | boy | girl |
| **Panel a: 2+ sample** | | | | | | |
| **IV: twins at second birth** | | | | | | |
| No. children | -0.077*** | -0.073*** | -0.096*** | -0.093*** | -0.098*** | -0.095*** |
| | (0.006) | (0.004) | (0.013) | (0.010) | (0.014) | (0.014) |
| N | 428547 | 428559 | 137443 | 137432 | 137277 | 137266 |
| **Panel b: 3+ sample** | | | | | | |
| **IV: twins at third birth** | | | | | | |
| No. children | -0.052*** | -0.053*** | -0.010 | -0.004 | -0.049 | -0.038 |
| | (0.009) | (0.009) | (0.026) | (0.025) | (0.032) | (0.031) |
| N | 208168 | 208146 | 33720 | 33706 | 33678 | 33664 |

Note: Robust standard errors are reported in parentheses and are clustered by household ID for 3+ sample.

Table 2.14: Robustness Check: Heterogeneity—Q-Q Trade-off by Gender of Older Sibling

| | Primary Enrolement | | Primary graduation | | Junior Enrolment | |
|---|---|---|---|---|---|---|
| | male | female | male | female | male | female |
| **Panel a: 2+ sample** | | | | | | |
| **IV: twins at second birth** | | | | | | |
| No. children | -0.023*** | -0.060*** | -0.023 | -0.040*** | -0.008 | -0.010 |
| | (0.006) | (0.005) | (0.024) | (0.015) | (0.014) | (0.018) |
| N | 204453 | 224667 | 68618 | 68942 | 68536 | 68858 |
| **Panel b: 3+ sample** | | | | | | |
| **IV: twins at third birth** | | | | | | |
| No. children | -0.008 | -0.035** | -0.093 | 0.010 | -0.100** | -0.009 |
| | (0.018) | (0.017) | (0.406) | (0.044) | (0.051) | (0.047) |
| N | 40416 | 80346 | 7076 | 10606 | 7058 | 10616 |

Note: Robust standard errors are reported in parentheses and are clustered by household ID for 3+ sample.

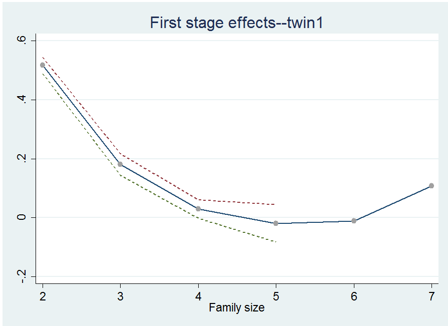
Table 2.15: Robustness Check: Heterogeneity—the Effect of Family Size on Education Outcome (1982 Census)

| | | Primary graduation | | | |
|---|---|---|---|---|---|
| | First stage | OLS | 2SLS[a] | NLIV[b] | GMM[c] |
| **Panel a: 2+ sample** | | | | | |
| **IV: twins at second birth** | | | | | |
| twin2 | 0.533*** | | | | |
| | (0.032) | | | | |
| No. childnum | | -0.023*** | 0.007 | -0.213*** | -0.041*** |
| | | **-0.025**[d] | **0.008** | **-0.236** | **-0.045** |
| | | (0.001) | (0.022) | (0.013) | (0.010) |
| Wald test | 280.160 | | | | |
| N | | 104487 | 104487 | 104487 | 104487 |
| **Panel b: 3+ sample** | | | | | |
| **IV: twins at third birth** | | | | | |
| **1st born** | | | | | |
| twin3 | 0.576*** | | | | |
| | (0.025) | | | | |
| No. childnum | | -0.032*** | 0.015 | -0.091*** | -0.019** |
| | | **-0.036** | **0.017** | **-0.102** | **-0.021** |
| | | (0.001) | (0.017) | (0.010) | (0.008) |
| Wald test | 531.440 | | | | |
| N | | 90233 | 90233 | 90233 | 90021 |
| **2nd born** | | | | | |
| twin3 | 0.605*** | | | | |
| | (0.035) | | | | |
| No. childnum | | -0.033*** | -0.005 | -0.130*** | -0.023* |
| | | **-0.037** | **-0.006** | **-0.146** | **-0.026** |
| | | (0.001) | (0.022) | (0.013) | (0.013) |
| Wald test | 299.29 | | | | |
| N | | 76734 | 76734 | 76734 | 76734 |

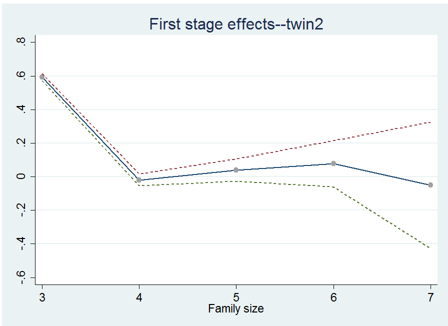
Note: Robust standard errors are reported in parentheses.
[a] 2SLS gives the two stage least square estimates. [b] NLIV gives the estimates based on non-linear first-stage. [c]GMM gives the result taking into account non-linear distribution in both first and second stage.
[d] Numbers in bold gives the corresponding transformation of the magnitude of Q-Q trade-off, i.e. **-0.025** means that an extra kid decreases the probability of primary enrolment by 2.5%, which corresponds to 2.3 percentage points decrease in the outcome variable compared to its mean value(0.023/0.904). The same calculation applies to all the estimates.

# Figures

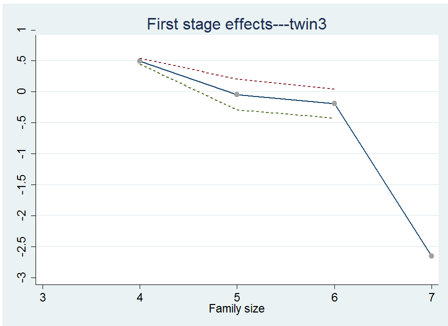

(a) Girl at first birth




(b) Twins at first birth




(c) Twins at second birth




(d) Twins at third birth

Figure 2.1: First Stage Effects on Fertility

Note: It plots the first stage effects of different instrumental variables on fertility, with number of children on x-axis and the first stage effects on y-axis. For example, the first dot in panel (a) tells that a girl at first birth increases the probability of fertility rises from 1 to 2 by 0.29. The dashed lines are estimated confidence intervals, note that at higher fertility range, the confidence interval is not shown due to the lose of significant power.

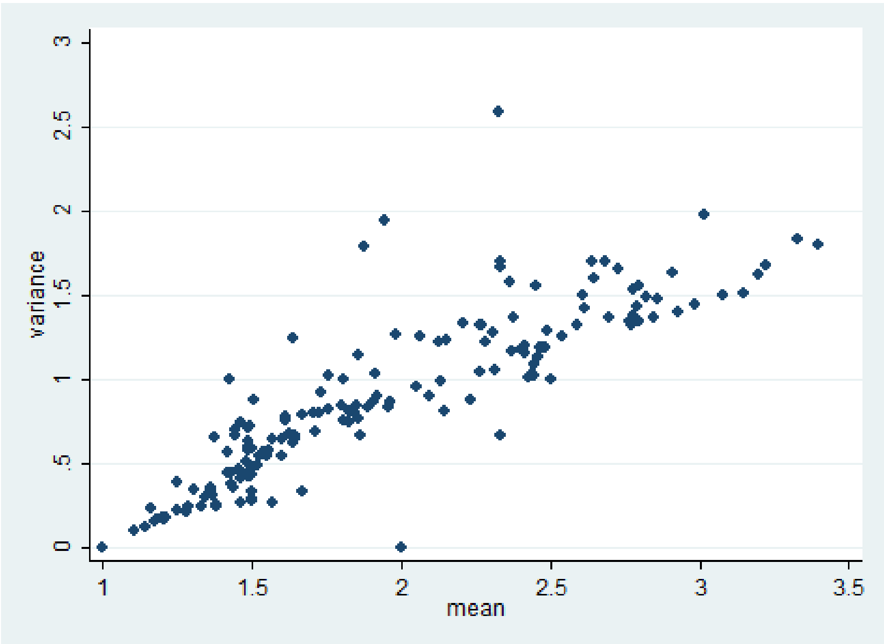
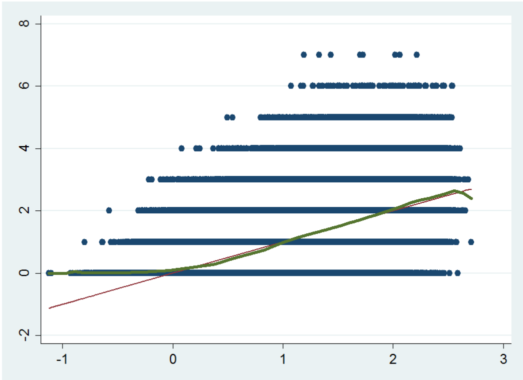
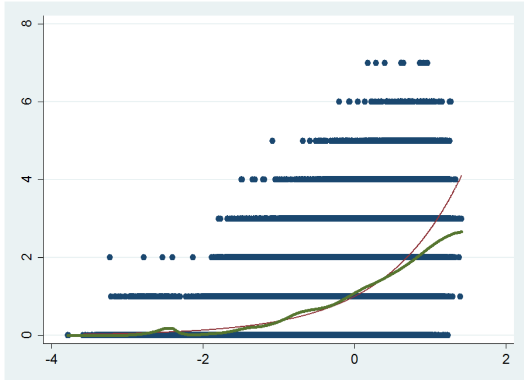



Figure 2.2: The Mean-variance Relationship for Families with Children

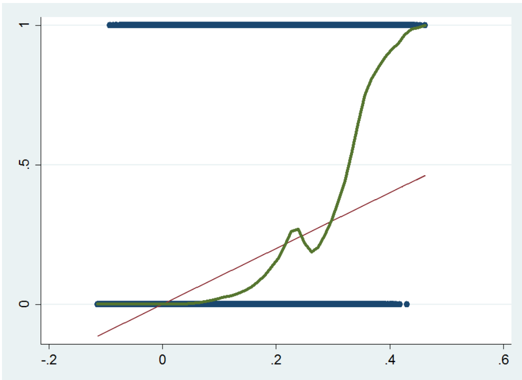
Note: It plots the variance against the mean conditional on the education level of mother and age of giving first birth.



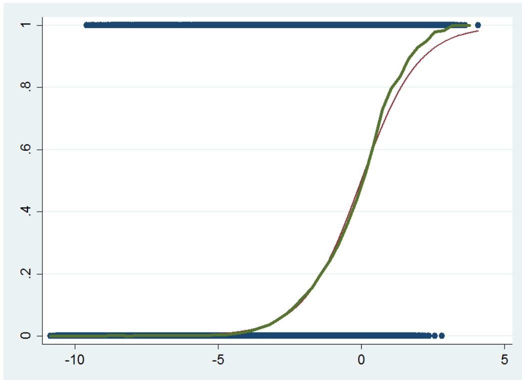

(a) first stage given by OLS




(b) first stage given by Poisson



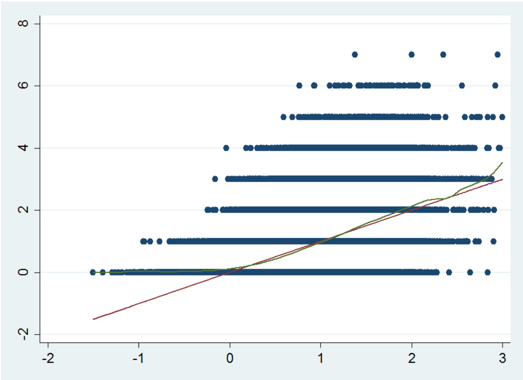

(c) second stage given by OLS



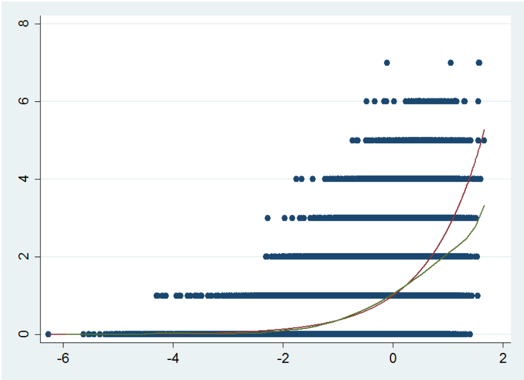

(d) second stage given by Logit

Figure 2.3: Model Fitting in Two Stages 1/2

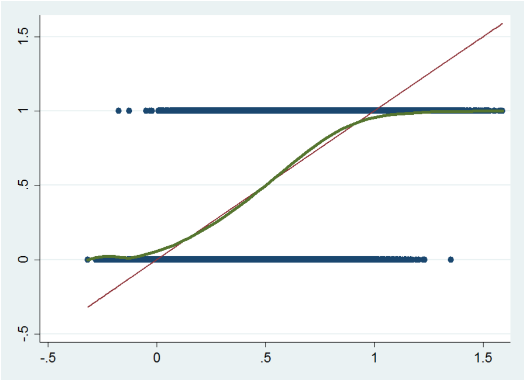
Note: Sample: at least 1 birth household; Endogenous variable: fertility; Outcome variable: the probability of completing primary school; Instrumental variable: gender of the first birth. The dots plot dependent variable $Yi$ or $Ni$ versus estimated linear indexes. The red line shows parametric fit of dependent variable $E(Yi|Xi)$ or $E(Ni|Xi)$ versus estimated linear indexes. The green thicker line represents a kernel regression of $Yi$ or $Ni$ on the fitted index for each model.



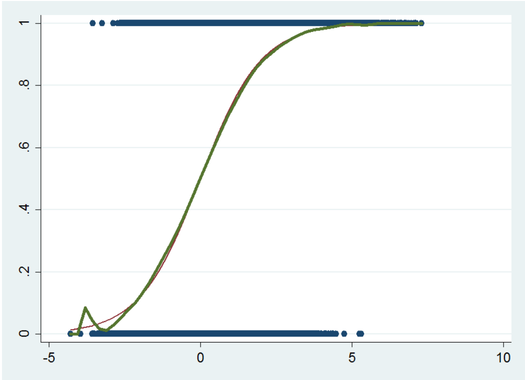

(a) first stage given by OLS




(b) first stage given by Poisson



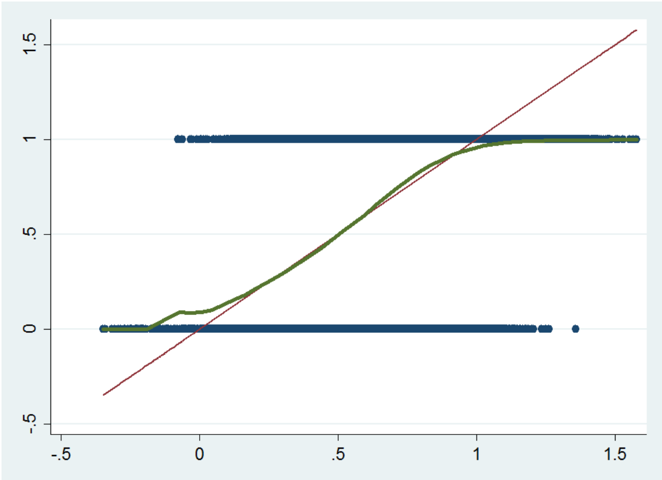

(c) second stage given by OLS



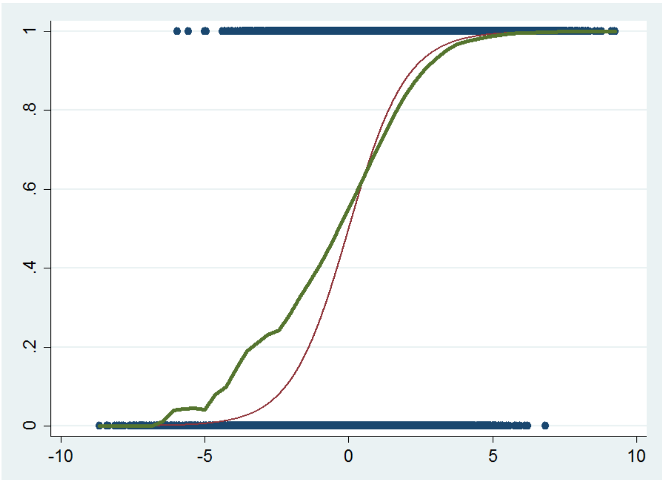

(d) second stage given by Logit

Figure 2.4: Model Fitting in Two Stages 2/2

Note: Sample: at least 2 birth households; Endogenous variable: fertility; Outcome variable: the probability of entering junior high school; Instrumental variable: twins at the second birth. The dots plot dependent variable $Yi$ or $Ni$ versus estimated linear indexes. The red line shows parametric fit of dependent variable $E(Yi|Xi)$ or $E(Ni|Xi)$ versus estimated linear indexes. The green thicker line represents a kernel regression of $Yi$ or $Ni$ on the fitted index for each model.



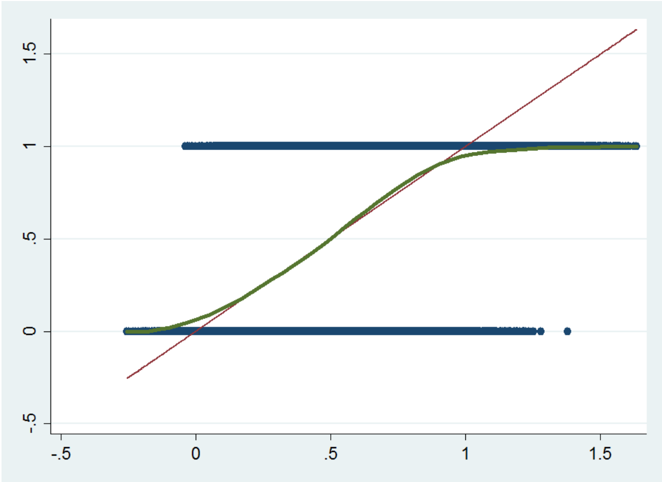

(a) 2SLS estimation



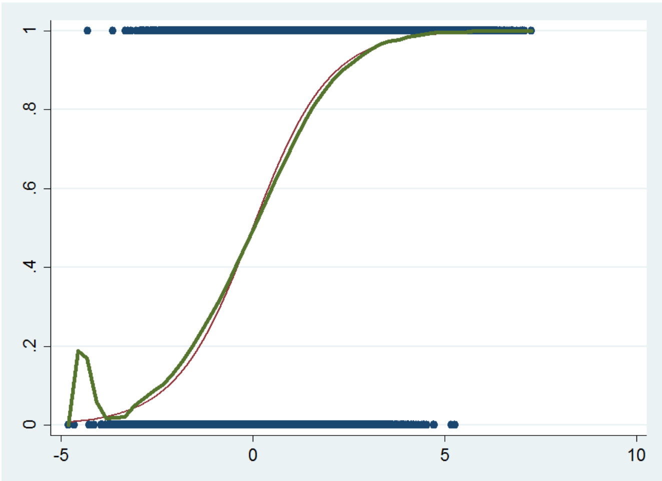

(b) GMM estimation

Figure 2.5: Model Fitting in One Step 1/2—2SLS VS. GMM

Note: Sample: at least 1 birth household; Endogenous variable: fertility; Outcome variable: the probability of completing primary school; Instrumental variable: gender of the first birth. The dots plot dependent variable $Yi$ versus estimated linear indexes. The red line shows parametric fit of dependent variable $E(Yi|Xi)$ versus estimated linear indexes. The green thicker line represents a kernel regression of $Yi$ on the fitted index for each model.




(a) 2SLS estimation




(b) GMM estimation

Figure 2.6: Model Fitting in One Step 2/2—2SLS VS. GMM

Note: Sample: at least 2 birth households; Endogenous variable: fertility; Outcome variable: the probability of entering junior high school; Instrumental variable: twins at the second birth. The dots plot dependent variable $Yi$ versus estimated linear indexes. The red line shows parametric fit of dependent variable $E(Yi|Xi)$ versus estimated linear indexes. The green thicker line represents a kernel regression of $Yi$ on the fitted index for each model.

# Chapter 3

## No Retirement Consumption Puzzle—the Effect of Labour Supply on Disaggregated Expenditures in the Later Life Cycle


### Abstract

This paper examines the life cycle profile of disaggregated nondurable expenditures for Chinese urban households. The result shows that aggregated expenditure conceals substantial heterogeneity in the life-cycle pattern of expenditure subcomponents. Based on the observed disaggregated life cycle pattern, I test how much of the adjustment of consumption across time and categories can be attributed to the labour supply status. A regression discontinuity approach shows that elderly households are able to maintain a stable consumption onset of retirement by adjusting expenditure across sub-aggregated categories and household behaviour. In particular, due to changes in family composition, family transfer, precautionary saving motives and adjustments in intra-household food production time, elderly households are not only able to smooth food expenditure but also food consumption upon retirement. Furthermore, by using subjective retirement expectations as an instrument for actual retirement, I find that the adjustments in sub-aggregated expenditure categories are resulting from unexpected retirement. This study confirms the prediction of Life Cycle Model and has important implications for using disaggregated data to test the existence of retirement consumption puzzle and for testing consumption theories.



**Key Words**: retirement consumption puzzle; sub-aggregated expenditures; regression discontinuity; subjective retirement expectations; China

**JEL Classification:**: D12, D91, J26, O12, P36

## 3.1 Introduction

The Chinese economy experienced rapid growth since the 1978 economic reform, which achieved an average of 8.6 percent in GDP annual growth over the period 1978-2007. The significant growth is mainly driven by the expansion of international trade. China has overtaken the US, becoming the world's biggest goods exporter in 2009, but this export-led pattern has proved to be unsustainable. The growth of the economy has decelerated recently, as signalled by the stock market crash in July 2015. Despite remarkable economic growth, household consumption, decreasing from 50% in 1980 to 36% in 2014, accounted for a gradually smaller proportion of the GDP. It raises the importance of analysing and understanding the consumption behaviour of Chinese households.

In addition, Chinese population has been ageing rapidly over the past three decades, mainly resulting from the family planning policy and longer life expectancy. The amount of elderly population was around 200 million by 2014, which accounted for 15.5 percent of the whole population, and it has been projected to surpass one third of the whole population by 2050. Given the fact that the working population is shrinking over time, the topic on delaying mandatory retirement age is widely debated. It is certainly a public policy that not only concerns economic growth, but more importantly the welfare of elderly people. Modigliani and Brumberg (1954) propose that households should maintain stable marginal utility over the life cycle, thus a failure to smooth consumption upon retirement would raise considerable concerns for the well-being of elderly people and adjustments of public policies. Hence, studying consumption behaviour of elderly individuals and their consumption response to retirement is of great policy relevance.

Studies of retirement consumption mainly focused on the western world where significant drops in consumption upon retirement are repeatedly documented. China provides a comparable context for investigating this issue due to different retirement policies and social norms. China has been implementing the mandatory retirement policy since 1950s. It firstly targeted at workers in public sectors and then gradually extends to private sectors since 1997. Although the mandated retirement age has been fixed for about 60 years, the replacement rate—the percentage of a worker's pre-retirement wage—varies across area, time and work-unit types. A major pension reform took place in 1997, which leads to the replacement rate of public sector workers decreased from 80% to 60% and the coverage extended to all types of enterprise[1]. However, the

[1]This includes state-owned enterprises, collective enterprises, foreign-funded enterprises,

implementation of the new pension system is far from achieving the policy goal. Li (2012) shows that the coverage rate increased from 24% in 1993 to 67% in 2011 for urban residents. China Household Income Project (CHIP) 2007 shows that 53.40 % of people residing in urban areas, age between 50 and 70, have no pension in 2007. Public sector workers typically receive higher replacement rate than workers in private sectors, this paper finds that it was 65% for the former and 50% for the latter across the whole sample period. The unique context of Chinese retirement and pension system provides a quasi-experimental framework for identifying a causal relationship between labour supply status and consumption. Moreover, elderly support prevails in Chinese culture, and the high saving rates of Chinese elderly households have been intensively reported(Modigliani and Cao, 2004; Wei and Zhang, 2011; Chamon and Prasad, 2010; Yang, 2012), which provides a unique opportunity to understand potential mechanisms through which households smooth their consumption.

This paper begins by exploring the life cyle profile of nondurable expenditure. In particular, it reports the evolution of household consumption at disaggregated categories over the life cycle. The importance of analysing subcomponents of consumption goods is raised by Aguiar and Hurst (2013) who claim that the life cycle pattern of total nondurable consumption is mainly driven by inputs that are complementary to work status and that are amenable to home production. This study confirms that total nondurable expenditure masks heterogeneity in the life cycle profile of disaggregated expenditures. Specifically, the total nondurable expenditure presents a typical hump-shaped pattern, which is consistent with the life cycle hypothesis with uncertainty. Food expenditure presents a hump-shaped profile in general but remains relatively stable at later stage of the life cycle. Core nondurable expenditure[1] shows a typical hump-shaped life cycle profile, while work-related expenditure[2] displays a significant drop since age 45. It seems that the pattern of total nondurable expenditure is mainly driven by the drop in work-related spending in the later stage of life cycle. These pieces of evidence emphasize the role of subaggregated expenditures in explaining the life cycle profile of total nondurable goods. Furthermore, this paper explicitly tests how retirement affects the consumption behaviour of households.

This study adds to prior literature in several aspects. First, it provides the

joint ventures enterprises, private enterprises, joint-stock enterprises and so on.

[1] Core nondurable expenditure includes rent or rent equivalent, utility, property management and daily use good spending. More detail will be provided in section 3.4.

[2] Work_related expenditure includes spending on clothes, transportation and communication. More detail will be provided in section 3.4.

first empirical evidence of the disaggregated Chinese household life cycle consumption profile. It shows that studies based on total nondurable expenditure ignore the heterogeneity in sub-aggregated expenditures. Second, it explicitly tests the life cycle model and presents new evidence that there is no retirement consumption puzzle in China. A regression discontinuity approach fails to find any impact of retirement on total nondurable expenditure, while retirement affects the allocation of expenditures on disaggregated categories. In particular, there is no drop in food consumption taking into account food production. Third, this study firstly applies subjective retirement expectation as an instrumental variable for the actual retirement in the context of China. I use an additional dataset (CHARLS) which allows me to separate anticipated changes in consumption from total changes in consumption upon retirement. Estimation result supports the life cycle model that adjustments in subcomponent consumption are resulting from unexpected retirement. Fourth, this paper differs from previous literature by exploring possible mechanisms through which the retirement affects disaggregated expenditures. Due to changes in family composition, increases in family transfers received from non-coresident children and stable household food production time, households are not only able to smooth food expenditure but also food consumption upon retirement. Work related expenditure drops upon retirement due to the complementary effects of labour supply, and it only significantly falls for high-educated households. Total nondurable expenditures stay stable leaving out sub-expenditures that are inputs into market work and amenable to home production, possibly also due to the high saving rates of elderly households. Fifth, this paper provides an implicit test of the consumption theory that the life cycle profile of consumption goods that are complementary to working status should be different from those goods that are substitution or amenable to home production. Estimation results confirm the complementary mechanism, and the home production theory finds support in adjustments of intra-household food shopping time. Sixth, the main analysis is based on a repeated cross sectional data that has the advantage over panel data in its longer span of coverage. Existing research in Chinese retirement consumption use panel structure with short time span which are insufficient to fully capture the welfare changes of elderly households.

The paper proceeds as follows: the next section presents institutional background and literature review on the analysis of household life cycle consumption behaviour. Section 3.3 states the empirical framework that the current paper focuses on. Section

3.4 describes the data source and descriptive statistics. Section 3.5 provides empirical results, robustness checks and possible smooth mechanisms of retirement consumption. Finally, Section 3.6 draws conclusion.

## 3.2 Background Information and Literature Review

### 3.2.1 Institutional Background

The mandatory retirement law has been officially implemented in public sectors[1] since May 1978, and then gradually applied to private sector. The retirement law requires male workers to retire at age 60, female white-collar workers to retire at age 55 and female blue-collar workers to retire at age 50, with few exceptions applying to certain occupations and physical situations. These age thresholds have not changed since the initial establishment in 1950s, while there are heated debates regarding postponing the retirement age given the fact that China is among those countries with earliest retirement age and experiences rapid changes in demographic structure. The ageing population puts heavy pressure on pension reforms that it has been projected that one pensioner will be sponsored by two tax payers in 2035.

The compliance of the mandatory retirement age law is not perfect, there are cases that people retire earlier than the mandatory age[2], and cases that people get re-hired by the previous employer after official retirement or get a new job.

Although the mandated retirement age has been fixed for about 60 years, the replacement rate varies across area, time and work-unit types. ChinaâĂŹs Urban EmployeeâĂŹs Pension System has gone through several reforms in changing the way providing funds and replacement rates. China established the urban pension system in the early 1950s, targeting at state-owned enterprise and collective enterprise workers, the latter made up a very small fraction of the total (Li, 2012). Under the planned economy system from 1950s to 1980s, this system covered all urban public sectors and

[1] Here public sector is a broad term, for the purpose of distinguishing from private sector. Public sector in this paper refers to governments and public sectors, state-owned enterprises, and collectively-owned enterprises

[2] The institutional reform took place in 1990s, aiming at reducing operation cost and reconstructing State-owned enterprises. Individual who is about to retire in 5 years and has worked more than 30 years can apply for an early retirement. Early retirees are able to receive certain amount of money from their employers monthly before getting pension upon normal retirement age.

the replacement rate was around 80%. With the reform and opening-up policy in 1980s, this system could not adapt to economic development and the government began to implement some local pilot social pension insurance programmes during the period 1991–1997. A major pension reform took place in 1997, which required the retirement fund to be provided by combining individual accounts and social pooling, rather than being entirely provided by employers before 1997. Under this context, the replacement rate decreased by 25% and the coverage rate increased from 24% in 1993 to 67% in 2011 for urban residents(Li, 2012).

This reform interferes with my sample period 1995-2007, but it should not affect the size of compliers and consumption behaviour right upon the retirement age given the nature of the Regression discontinuity design. Additionally, results are robust to drop 1995 wave.

### 3.2.2 Literature Review

Aguiar and Hurst (2013) firstly document that the life cycle profile of total nondurable consumption masks surprising heterogeneity across expenditure subcomponents. The well-known hump-shaped nondurable expenditure profile is mostly driven by sub-expenditures that are inputs into market work or are amenable to home production. Consumption categories that are complements to labour supply would decline as the labour supply declines at the later life cycle, while those categories that are substitutes to labour supply would not decline significantly with the change in labour supply since they are cost opportunity of the working status.

Their paper gives insight to a new understanding of the so-called retirement consumption puzzle—that household consumption drops significantly at retirement—which contradicts the life cycle hypothesis by Modigliani and Brumberg (1954). A key implication of the life cycle model is that individuals are able to smooth their consumption over the life cycle against expected income shocks. While retirement is probably the most predictable change in income, Banks et al. (1998) firstly document the retirement consumption puzzle in the UK, they claim that unexpected adverse information around retirement would affect the consumption behaviour. Smith (2006) and Barrett and Brzozowski (2012) show that it is the involuntary retirement that causes a drop in the consumption while the voluntary retirement actually shows no change. Bernheim et al. (2001) find a pronounced discontinuity in consumption at retirement by using Panel Study of Income Dynamics (PSID) in US, in particular, they claim that work-related expenses or leisure substitutes do not contribute to explain the reduction in consumption based on Consumer Expenditure Survey (CEX). However, based on the longitudinal component of CEX, Aguila et al. (2011) find no evidence of discontinuity of retirement consumption measured by total nondurable expenditure, while food expenditure declines by 6 percent at retirement.

Hurst(2008) surveys the retirement consumption puzzle literature and concludes that the standard life cycle consumption model augmented with home production and health shocks accounts for household retirement consumption, to a large extent. The fact that work-related expenditure declines at retirement has been consistently documented but what really puzzling is that the well documented significant drop in food expenditure at retirement. Given food is necessity in daily life thus have relatively lower income elasticity, a promising explanation for the expenditure drop is the relative lower opportunity cost of home production after retirement.

Empirical challenges to identify a causal effect of retirement on consumption come from several aspects. The first concern is that retirement is an endogenous choice, but often times previous studies have no good ways to deal with the endogenous issue, therefore, it is most likely for them to present a correlation instead of a causal effect. There are few exceptions that exploit variation in retirement due to retirement benefit eligibility (Battistin and Weber, 2009) or subjective retirement expectations (Haider and Stephens, 2007).

Battistin and Weber (2009) identify the causal effect of retirement on consumption in Italy by using the exogenous variation in pension eligibility. They find that a 9.8 percent of the nondurable consumption drop is due to eligibility induced retirement, and the drop is explained by reduction in work-related expenses or leisure substitutes. However, most of the retirement consumption drop is attributed to a significant reduction in the number of adult children living with their parents.

The paper most closely related to this study in the context of China is provided by Li et al. (2015). They exploit a causal relationship between retirement and consumption within a regression discontinuity framework. The identification comes from China's mandatory retirement policy and they report a 21 percent drop in nondurable consumption at retirement, but the drop is mainly driven by food at home and work-related expenditures. Their work is in support of the finding of Hurst (2008) that the retirement consumption puzzle does not exist when the life-cycle model is augmented with home production. Dong and Yang (2016) find a significant decline in food expenditure that is mostly explained by drop in average food prices rather than quantities, but find no changes in work-related expenses. This paper differs from these two studies in that new evidence on food expenditure is provided, more importantly several mechanisms through which elderly households are able to smooth consumption are proposed and tested. Furthermore, subjective retirement expectation is employed as an instrument for better understanding the consumption smooth channel.

The second challenge is that most of the literature have no detailed information on consumption or they provide evidence based only on food expenditure or total nondurable expenditure. While food is a relatively easy accessed measurement of expenditure in most of the database and assumed to have low income elasticity, it cannot represent a composite measurement of expenditure especially in the presence of heterogeneity in the subcomponent consumption.

## 3.3 Empirical Methodology

To provide the life cycle profile of mean expenditures, this analysis adopts the specification by Aguiar and Hurst(2013),

$$lnC_{it}^k = \alpha_0^k + \alpha_1^k Age_{it} + \alpha_2^k Cohort_{it} + \alpha_3^k Family_{it} + \epsilon_{it}^k \tag{3.1}$$

where $C_{it}^k$ represents nondurable expenditure of household $i$ in year $t$ on category $k$, $Age_{it}$ contains a set of age dummies that cover household head age ranging from 26 to 75, $\alpha_1^k$ shows the effect of age at each life stage relative to age 25 on expenditure category $k$. $Cohort_{it}$ is a vector of cohort dummies, ranging from 1927 to 1974. $Family_{it}$ is a set of family composition dummies, including dummies indicating number of household members, dummies representing number of children in age categories 0-2, 3-5, 6-13, 14-17 and 18-21 and a dummy for marital status. The measure of expenditure is at household level instead of the individual level, hence I use the cross-sectional differences in family structure to identify family composition effects. One common alternative approach is to deflate expenditure by a measure of common equivalence scales, as this study looks at sub-aggregated expenditure categories, the major limitation in applying a common equivalence scale is that it cannot account for the different returns to scale across expenditure categories(Aguiar and Hurst, 2013).

Since this analysis is based on repeated cross section surveys, the age effects would actually capture a mixture of life cycle and cohort effects. Hall(1968) states that age effects are not identified due to the collinearity among age, cohort and year effects. The standard practice is to normalize year dummies so that the year effects are capturing cyclical fluctuations or business cycle effects that average to zero over the long run (Deaton, 1997). Thus the year effects would be orthogonal to a time trend. This study deals with few waves that it is insufficient to isolate time trends from transitory shocks. Moreover, if the true time effects contains a linear trend, the orthogonal assumption will create bias by forcing that trend into estimated age and cohort effects (Browning et al., 2014). An alternative to put restriction on the time effects is to model time effects with observable variables. To account for the relative change in the price of goods over time, expenditures are deflated based on consumption category, region of residence and year.

To test how the consumption behaviour of elderly people changes with working status, specifically to assess the impact of retirement on sub-aggregated expenditures, this paper exploits the retirement consumption under a regression discontinuity (RD) framework. The RD design was firstly introduced by Thistlethwaite and Campbell (1960) and become gradually popular in relatively recent studies(Angrist and Lavy (1999), Hahn et al. (2001), Imbens and Lemieux (2008), Lee and Lemieux (2010)). There are few studies employing the RD design to test the existence of a retirement consumption puzzle(Battistin and Weber (2009), Stancanelli and Van Soest (2012), Li et al. (2015), Dong and Yang (2016)). Essentially, I would like to estimate the treatment effect at a range as close as possible to the threshold of assignment variable while not losing too much precision around the cutoff.

The fact that retirement is mandatory at a certain threshold in China provides a unique quasi-experiment to capture the causal effect of retirement on consumption. Under the assumption that the cutoff is not related to any family characteristics or predetermined variable, any significant change in consumption at the age cutoff can be attributed to the retirement.

Ideally this paper is interested in estimating the following specification:

$$lnC_{it}^{k} = \gamma_0^k + f(age_{it}) + \rho Retire_{it} + \nu_{it}^{k} \tag{3.2}$$

where $f(age_{it})$ represents a nonlinear relationship between outcome and the forcing variable $age_{it}$ for allowing a reasonable smooth at both sides of the cutoff. In practice, there are two ways to approximate $f(age_{it})$, one is to use a $pth$ order polynomial specification i.e.

$$lnC_{it}^k = \gamma_0^k + \gamma_1^k \tilde{age}_{it} + \gamma_2^k \tilde{age}_{it}^2 + ... + \gamma_p^k \tilde{age}_{it}^p + \rho Retire_{it} + \nu_{it}^k \tag{3.3}$$

where $\tilde{age}_{it} = age_{it} - 60$, centering the running variable at cutoff age 60 to ensure that the treatment effect at $age_{it} = 60$ is the coefficient on $Retire_{it}$ in a regression with interaction terms(equation 3.5 – 3.8). The choice of polynomial order is based on the Akaike information criterion (AIC) as suggested by Van der Klaauw (2002) and Lee and Lemieux (2010). In the regression context, the AIC is given by

$$AIC = Nln(\sigma^2) + 2k \tag{3.4}$$

where $\sigma$ is the root mean squared error, and $k$ represents the number of parameters in the regression.

$Retire_{it}$ is a dummy for retirement status of the household head in household $i$ at time $t$. This specification implicitly assumes that the only factor affecting the outcome at a range very close to age 60 is the retirement status which is independent of the error term. Figure 3.2(a) shows that there is a jump of around 40 percent of individual transiting into retirement from employment. However, there are non-compliers who got early retirement or postponed retirement. Hence the endogeneity of retirement would bias the result estimated from equation 3.3. Following Li et al.(2015), I use an age indicator $D_{it}$ that equals to one if age above 60 and 0 if below 60 to instrument the retirement status. $D_{it}$ is a strong predictor for retirement and is very unlikely to suffer from any manipulation and selection bias.

The first stage specification can be estimated from

$$\begin{aligned} Retire_{it} =& \delta_0 + \delta_{l1} \tilde{age}_{it} + \delta_{l2} \tilde{age}_{it}^2 + ... + \delta_{lp} \tilde{age}_{it}^p + \lambda D_{it} + \\ & \delta_{r1} \tilde{age}_{it} D_{it} + \delta_{r2} \tilde{age}_{it}^2 D_{it} + ... + \delta_{rp} \tilde{age}_{it}^p D_{it} + e_{it} \end{aligned} \tag{3.5}$$

where both $D_{it}$ and the interaction terms are used as instruments for $Retire_{it}$.

The second stage model with interaction terms can be specified as follows:

$$\begin{aligned} lnC_{it}^k =& \gamma_0^k + \gamma_{l1}^k \tilde{age}_{it} + \gamma_{l2}^k \tilde{age}_{it}^2 + ... + \gamma_{lp}^k \tilde{age}_{it}^p + \rho Retire_{it} + \\ & \gamma_{r1}^k \tilde{age}_{it} Retire_{it} + \gamma_{r2}^k \tilde{age}_{it}^2 Retire_{it} + ... + \gamma_{rp}^k \tilde{age}_{it}^p Retire_{it} + \nu_{it}^k \end{aligned} \tag{3.6}$$

A natural alternative to approximate $f(age_{it})$ is to use a nonparametric kernel method for showing a local linear fit. By applying kernel weights and conditioning on the age, the idea is to eliminate residual differences between individual who are close to get retired and are just retired. The first-stage and reduced-form local linear kernel regressions are specified as follows,

$$\min_{\alpha,\delta} \Sigma_{it} K(\frac{\tilde{age}_{it}}{h})(Retire_{it} - \alpha D_{it} - \delta_0 - \delta_1 \tilde{age}_{it} - \delta_2 \tilde{age}_{it} D_{it})^2 \tag{3.7}$$

$$\min_{\beta,\gamma} \Sigma_{it} K(\frac{\tilde{age}_{it}}{h})(lnC_{it}^k - \beta D_{it} - \gamma_0 - \gamma_1 \tilde{age}_{it} - \gamma_2 \tilde{age}_{it} D_{it})^2 \tag{3.8}$$

where $K(\frac{\tilde{age}_{it}}{h})$ is a triangular kernel that puts more weight on observations that closer to the cutoff point, $h$ is a positive bandwidth sequence. Here the running variable comes with discrete nature so that it simplifies the problem of bandwidth choice, as the graphic fit will be shown using the mean of the outcome variable for each value of the discrete running variable.

Estimated effects should be interpreted as a local average treatment effect since fuzzy RD by construction is a 2SLS estimation with treatment-covariate interactions. The estimated causal effect captures the average treatment effect for those compliers who only retire at the point when they reach the mandatory retirement age.

## 3.4 Data

The empirical analysis is based on data from the urban Chinese Household Income Project (CHIP), which was conducted by the Institute of Economics at the Chinese Academy of Social Sciences with assistance from the National Bureau of Statistics(NBS). The survey adopts a national probability sample of households approach, selecting provinces from four distinct regions varies in economic development and geography for constructing a nationally representative sample. The four regions of China are defined as coastal, central, western and large municipalities with provincial status.

I use four waves out of the available five from this project, which covers year 1995, 1999, 2002 and 2007.

The initial 1988 survey differs from the rest in terms of purchasing pattern. In order to meet people's need of necessities against a terrible shortage of supply since 1950s, the government implemented a central controlled food supply policy that urban people were issued with food ration which was the only way to purchase food. This policy was abolished nationwide in 1992. One may suspect that the cease of this policy would induce a change in people's expenditure pattern, thus the 1988 wave is excluded.

All surveys document individual and household characteristics, sources of income and categories of expenditure. All expenditure and income data are deflated on urban and category basis, and are in 2007 constant Chinese Yuan[1].

The expenditure data are recorded on a diary basis across the survey years in the CHIP dataset. There are three main disaggregated expenditures to be exploited in this study. The first category is food expenditure, the second category is work-related expenditure which includes clothing, transportation and communication. The third category is the core nondurable expenditure which includes rent or rent equivalent, utilities, property management and daily use goods spending. This study only focuses on nondurable expenditure since it is difficult to obtain annual service flow measures from durables. For the same reason, education and health expenditure are excluded.

There is no information for distinguishing food at home from food away in the sample, except for 2002 urban survey. Hence, food expenditure consists of both food at home and food away from home. It is also difficult to distinguish work-related transportation from travel-based transportation spending in the sample. Given the fact that for most of the working age people, travelling transportation should account for a relatively smaller fraction of total transportation cost in China during the sample period, this study assumes that most of the transportation cost occurred during working age is driven by work-related transportation cost. The drop in transportation cost at later stage of life potentially confirms this statement. In addition, the robustness check section uses a demand system analysis to test the relationship between work-related spending and work status by directly controlling for labour supply. Utilities costs include electricity and fuel expenditures. Entertainment expenditure is included in a broader category—education, culture and entertainment services. For most of the survey waves, it is difficult to isolate the entertainment expenditure from the rest. Thus entertainment expenditure is not included in our analysis. Rents are computed as annual actual rent paid of renters and self-reported estimation of rental equivalent of the house owners.

The sample is restricted to households who report non-missing expenditures on five key subcomponents of the consumption: food, transportation and communication, clothing, utilities and rent or equivalent. This condition is not overly restrictive, leading to a reduction in sample size by 2 percent. The sample only includes households with household heads age over 25 and below 75 to ensure that household head has completed education. The pooled sample left with 21952 observations. This sample is used to

[1]Results are robust to deflating based on CPI index by province and category.

describe the life cycle consumption pattern of households aged from 25 to 75. It will be restricted further for the testing of retirement consumption puzzle.

Table 3.1 shows the share and mean change of three main nondurable goods over the life cycle. Specifically, it shows the pattern for individual ages around 25, 45 and 65. The share of food spending over total nondurable expenditure increases across these age groups, while work-related expenditure share decreases and the core nondurable share remains fairly constant. Column (4) and (5) show the log change in mean expenditure between age 25 and 45, 45 and 65, respectively. The log change is given by the change in corresponding age dummy coefficient as shown in figure 3.1.

In this analysis, retirement is defined based on current labour force status, i.e. individual is retired if self-reported labour supply status is retired. For testing the existence of retirement consumption puzzle, the RD sample restricts households to those with male household head and age between 50 to 70 (60 is excluded[1]). This restriction further shrinks the sample size, the RD sample contains 5160 observations. This analysis focus only on households with male household heads due to the fact that the mandatory retirement law is more complicated for women, depending on the occupation and type of work unit. In addition, there are few cases that female are the household heads. In fact, the average age difference between husband and wife is around three years old, 71.5(95.4) percent of the households in our sample have less than 5(10) years age-gap between husband and wife, which suggests that most of the spouse are already retired, if spouse was employed, upon husband's retirement. Thus, this analysis does not take into account wife's retirement decision and attributes the change of household consumption at husband age 60 mainly to the retirement of husband.

Table 3.2 shows the summary statistics for both outcome variables and some predetermined variables of the RD sample. The mean age of household head is around 58, 4% of which are ethnic minorities, with 41% obtained high school or above education. On average, there are 3.05 individuals in the household among which includes 0.79 child. Around 98 percent of the male are married or cohabiting, with 63 percent of the spouse are also retired. The majority(88%) of household head was/is employed in the public sector, with 11% were/are employed in government, 16% were/are employed in institutions, 54% were/are employed in state-owned enterprises or collective-owned enterprises, 19% were/are employed in private sectors[2]. Table 3.2 also reports the mean

[1]Individuals age 60 are likely to be in a situation of a mixture of pre-retirement and post-retirement status, therefore the expenditure data of age 60 people are excluded.

[2]This paper defines two broad categories of workunit: public and private sectors. Public sector includes government organizations, institutions, state-owned and collective-owned

expenditures in Chinese currency[1].

## 3.5 Results

This section starts with showing empirical life cycle patterns of the evolution of mean expenditures. Based on the graphic evidence, I test the association between labour supply status and consumption goods. To be specific, the RD approach is employed to test how much of the drop in total nondurables and changes in the three sub-aggregated expenditures in later stage of the life can be attributed to transitioning into retirement.

### 3.5.1 Empirical Patterns

Figure 3.1 shows the life cycle consumption pattern of total nondurable goods and subcomponent expenditures. The total nondurable expenditure seems to display a hump-shaped pattern but with two peaks, it rises over the early stage of life cycle, peaking at around age 40 at roughly 28 log points higher than the baseline 25-year-old spending, the second peak comes at age 55 with 26 log points higher than the baseline age effects[2]. The consumption pattern documented in figure 3.1(a) conceals substantial heterogeneity among subcomponent consumption goods as depicted in figure 3.1(b). It shows that different expenditure categories present quite heterogeneous life cycle patterns for both mean and variance of spending. In general, both food and core nondurable spendings show a hump-shaped profile. Food expenditure rises by 30 log points at around age 40 and then stays relatively stable until age 54, followed by some fluctuations within the range between 20 and 40 log points. Core nondurable expenditure presents an increasing pattern before age 35 and stays relatively stable between age 35 and age 50, then it slightly declines. Work-related consumption displays a dramatically different life cycle profile from the previous two consumption goods. It drops substantially after around age 45 which is below the baseline age effect at age 25, followed by a decline of about 60 log points until age 75[3].

In summary, there is substantial heterogeneity across life cycle profiles of expenditure categories, which is consistent with the finding of Aguiar and Hurst(2013). However,

enterprises. Private sector includes private firms, sino-foreign joint ventures, foreign companies, state share-holding companies, and other ownership.

[1] Here the expenditures are measured in Yuan for the ease of understanding the composition of spending, but main results are reported based on log value of expenditure.

[2] Note that Aguiar and Hurst (2013) show a peak at age 45.

[3] Aguiar and Hurst (2013) show a roughly 60 log points fall after middle age.

in particular, I do not observe a fall in mean food expenditure at retirement as mostly documented in the literature. It is natural to incorporate the home production theory since the opportunity cost of time is relatively lower after retirement. If this also applies to this study, then the actual food consumption might increase after retirement. As the observed pattern has shown that work-related spending drives down the total nondurable expenditure at later stage of the life, a potential explanation for the heterogeneous pattern can be derived from the relationship between different expenditure categories and employment status. Those categories which are complementary to work should fall while those substitutes should increase.

### 3.5.2 Main Results

Given the potential importance of work-related expenditures in accounting for the evolution of the life cycle profile, this paper is interested in assessing the relationship between labour supply and adjustment of expenditures across life cycle and expenditure categories. Specifically, in order to understand how much of the decline in spending of the elderly households(age between 50 and 70) can be attributed to transition into retirement, a RD design is employed to test the existence of retirement consumption puzzle in China. The share of work-related spending accounts for around 20 percent of the total expenditure at age 50-70, clearly not a negligible amount. It is natural to question how much of the total nondurable drop at around retirement age can be attributed to a change in work status. Here, the real consumption puzzle would present if the total nondurable consumption drops at retirement even after taking out the work-related spending.

Figure 3.2(a) shows a significant jump of the retiree fraction at the mandatory retirement age 60, with the share of retirees increasing from around 40 percent at age 59 to around 81 percent at age 61. Table 3.3 confirms a significant first stage effect that age above 60 increases the likelihood of retirement by about 35.0% to 46.5%, depending on different specifications used. As stated in equation 3.5, this study incorporates flexible age controls, allowing for interactions between age dummies and $D_{it}$ and asymmetric polynomial orders on either side of the age 60 cutoff. A valid RD design would imply that the inclusion of other covariates does not affect the consistency but the efficiency of point estimates. Table 3.3 shows that province and year dummies attributes to explain the variation in the outcome variable retirement, and the standard errors of augmented regressions are smaller while the significance and magnitude of the estimators stay quite similar.

Tests of validity of the RD design are presented in table 3.4, additionally in figure 3.2(b) and figure 3.3. In order for the RD design to be valid, one would expect individuals to have imprecise control over the assignment variable. Since it is not possible to test this assumption directly, an alternative approach is to test the continuity of the density of assignment variable at age 60 (McCrary, 2008). If individuals near the compulsory retirement age were able to select on either side of it, and if they did so in a monotonic way, then one would expect to observe a discontinuity of the density of running variable at age 60[1]. Figure 3.2(b) provides evidence that retirement status is as good as randomly assigned near the cutoff 60.

Ideally, a discontinuity in the outcome of interest should only come from the treatment effects while there should be no correlation between predetermined variables and the treatment. The covariables that are taking into consideration are household size, number of children and number of children in finer age groups 0-2, 3-5, 6-13, 14-17, 18-21, education level, ethnic minority, whether the household head is married or cohabited, the retirement status of his spouse. Figure 3.3 tests the validity of the assumption that the cutoff is not correlated with any predetermined background variables[2]. Table 3.4 also shows that there is no correlation between the treatment and household background characteristics, which provides additional support on the fundamental assumption that retirement status is as good as randomly assigned to either side of the cutoff controlling for a smooth polynomial in age. Here the polynomial order is chosen by the AIC criterion. Since the AIC is a generalized cross validation procedure that minimizes a loss function, it puts more weight on model fitting rather than a significance testing of the null hypothesis that the estimator is zero. It explains why the AIC procedure selects specification with polynomial orders that lack of significance or nearly zero in magnitude.

The main result of the regression discontinuity design is given by figure 3.4 and table 3.5. Figure 3.4 shows a discontinuity in work-related expenditure, while there is no clear evidence of a drop in total nondurable expenditure, food and core nondurable

[1]Most likely age is considered as impossible to be manipulated. Out of caution, this paper still conducts this density test to exclude the possibility for people to deliberately change age. Additionally, the age is documented only in years not in months in some of the survey waves. This density tests helps to check the problem of age rounding in the case of self-reported age.

[2]There is an observed discontinuity in the proportion of minority households. It is hard to connect this pattern with any events or policy changes related to minority retirement. Given that the minority households account for only 4 percent of the whole sample, the drop is possibly driven by very few observations with characteristics that are uncorrelated with retirement decision.

expenditure. Table 3.5 confirms the graphic finding that there is no significant impact of retirement on total nondurable expenditure; food and core nondurable expenditures stay stable, while work-related spending shows a significant decline of 42.4%. It implies that households are able to adjust expenditures across categories even though there is no significant change in total nondurable spending. Hence, households smooth consumption through reallocating sub-expenditures.

Note that the treatment-determining covariate is a discrete variable, since it is impossible to get observations within arbitrarily small neighbourhoods of the cutoff, the identification relies on choosing a particular functional form for relating the covariate to outcomes. The specification error occurs whenever there is a deviation of the expected value of outcome from the predicted outcome of a given functional form. This study assumes that the specification error is random and identical (Lee and Card, 2008), to be specific, I assume that the sources of specification error is independent of treatment status. This orthogonal assumption of the specification error may seem restrictive, but to an extent it relaxes the assumption of no specification error that conventional inference relies on. The randomness assumption induces a within-group correlation (at age level) in the error, therefore consistent estimators are provided by clustering at age and province level.

### 3.5.3 Robustness Check

Several robustness checks have been applied to this analysis, mainly focusing on testing the robustness of the regression discontinuity method and the sensitivity of the expenditure measurements.

#### 3.5.3.1 Bandwidth Choice

The discrete nature of the running variable simplifies the problem of bandwidth choice, each observation in the graphic presentation displays an average over each age. The main analysis focus on observations with ten-year bands on each side of the mandatory retirement age, for achieving a bias-efficiency balance.

Table 3.6 experiments with different selection criteria on the bandwidth, ranging from five-year bands to fifteen-year bands. Results are robust using the eight-year up to fifteen-year bands that a significant decline in work-related expenditure is observed upon retirement, while seven-year bands and smaller bands show similar drops in magnitude as observed in table 3.5 but insignificant. Figure 3.1(b) shows that work-related expenditure starts to drop at around age 45, wider bandwidth captures a bigger work-related spending drop and composition effects due to mortality. Although conventionally small bandwidth reduce the bias, in our case smaller bands suffer from the effect of outliers in the age profile. Thus it provides less credibility on the estimated effects upon retirement.

#### 3.5.3.2 Alternative Specifications

Figure 3.2, figure 3.3 and figure 3.4 are based on local linear fit estimations, which provide a robustness check on the AIC-selected polynomial specifications as shown in table 3.3, table 3.4 and table 3.5. Both the local linear and AIC polynomial specifications reach the same conclusion.

The main analysis draws conclusion based on specification chosen by the AIC criterion, which incorporates third order polynomial in some cases. However, higher order specification achieves more flexible control and reduces bias at the cost of greater asymptotic variance. Gelman and Imbens (2014) argue that regression discontinuity analysis should not control for higher order (third, fourth, or higher) polynomials of the running variable. The estimators that rely on such methods have poor properties and are often misleading, therefore instead they suggest to use quadratic polynomials or other smooth functions.

Figure 3.5, figure 3.6 and figure 3.7 show graphic results that are based on quadratic polynomial regression, and table 3.7 shows the main regression of quadratic specification. Results confirm the robustness of previous findings, but one exception is that work-related expenditure shows no discontinuity under this specification. Although the coefficient gives similar magnitude as the AIC selected model, the standard error of work related expenditure doubles compares to the AIC selected model.

#### 3.5.3.3 Placebo Test

Essentially, the placebo test in this context is to estimate whether there are discontinuities at points where there should be no jumps. This analysis implements this idea by splitting the sample into two subsamples on each side of the cutoff value and then take the median of the running variable as a “fake” cutoff in each subsample (Imbens and Lemieux, 2008). If the RD design is valid, one would expect no discontinuity exists in either subsample. Picking the “fake” cutoff at the median on either side increases the power of the test to detect jumps. Moreover, this idea avoids including the true cutoff point in subsamples, therefore excludes the possibility that the known discontinuity contaminates the estimation.

Placebo estimates are implemented with age 54 and age 64 as the “fake” cutoff points on either side of the true cutoff. Table 3.8 is in favour of the null hypothesis of no effect of retirement on either side of the cut-off. Notice that the standard error in both subsamples are relatively bigger due to smaller sample size.

#### 3.5.3.4 Heterogeneity

Households with different pre-retirement wealth are likely to respond in different ways to retirement even if the shock is expected. The socially disadvantaged households maybe less prepared for the reduction in income, such as a shortage of saving or less organised financial behaviour or bounded by credit constraint. Due to a lack of wealth measurement across all waves, education is used as a proxy for wealth. Table 3.9 shows that there are no significant differences in the consumption response to retirement between low and high educated groups. One exception is in the aspect of work-related expenditure, high educated households show a significant 37.5 percent of a decline while low educated households show similar magnitude of effect, not significant though. Further tests suggest that the significant drop among high educated group is resulting from a significant decrease in clothing of 49.2 percent[1]. Potential explanations could be that high educated households are more likely to be employed in certain occupations or work-units that stress more importance on outfits, in contrast, low educated households incline to pay less attention on the outfits.

Chinese pension system is renowned for its double standard treatments, workers in government, institutions and state-owned enterprise enjoy higher replacement rates than those work for private enterprises. There are different measurements of replacement rate, ideally it represents the rate of pension and pre-retirement wage for the retirees. Unfortunately the data does not provide information on both pre-retirement wage and pension for the same individual. Alternatively, this paper obtains the average replacement rate by dividing the average pension of retired individuals by average wage of the working force across cities. The replacement is around 50% among private sectors and 65% among public sectors. One would expect that workers in former sector are less affected by retirement than the latter. An attempt to exploit the consumption retirement puzzle through different work units is presented in table 3.10. However, due

[1]This estimation is not presented in this paper but is available upon request

to the few observations of private sector, the estimates are not precise. Additionally, this paper considers the heterogeneity of retirement effects across time. Table 3.11 shows no significant drop in total nondurable expenditure across years and that the biggest drop in work related expenditure occurred in 1995. Similarly, the estimates are unfortunately imprecise.

#### 3.5.3.5 Measurement of Expenditure

Given the fact that clothing, transportation and communication are categorized as work-related expenditure, it is natural to question whether it picks up some expenditures that are not related to work. For example, transportation expenditure includes not only commuting to work but also travelling and visiting family members and friends. Since the author has no sources of data to tease out that part of expenditure, a potential approach is to directly measure the effect of work status in the life cycle equations. A demand system analysis (Deaton and Muellbauer, 1980; Aguiar and Hurst, 2013) has been employed to assess how much of the evolution of subaggregated expenditures can be attributed to the labour supply status.

A main concern with this regression is that the labour supply is closely related to permanent income which makes the specification suffer from omitted variable bias. For example, if labour supply is measured by employment status dummy, then those individuals who are working are likely to earn more than those who are unemployed. By conditional on total expenditure which is a proxy for permanent income, it is able to separate the effect of labour supply from income effect. There are some other issues with this specification: first, total spending appears as both a right hand side variable and a left hand side dependent variable, which makes it subject to measurement error. Specifically, any measurement error found in one subcomponent of expenditure will contaminate the total expenditure $X_{it}$. The standard practice is to instrument total nondurable expenditure with income and education. Second, labour supply is potentially endogenous if it is correlated with some shocks that affecting the expenditure share. For example, if there is a shock on transportation industry that makes transportation more expensive and therefore tends to be a larger share of total spending(assume most part of the transportation cost is unavoidable), in the meanwhile this shock affects the employment status of people working in transportation industry, then failing to instrument labour supply would lead to a bias. Moreover, the instruments used for controlling measurement error would be vulnerable to endogeneity issues as labour supply suffers.

$$Cshare_{it}^{k} = \theta_0^k + \theta_1^k Age_{it} + \theta_2^k Cohort_{it} + \theta_3^k Family_{it} + \theta_4 lnX_{it} + \theta_5 L_{it} + u_{it}^k \quad (3.9)$$

where $X_{it}$ represents the total nondurable expenditure including food, work-related and core nondurables; $Cshare_{it}^{k}$ measures the share of expenditure category $k$ out of total nondurable spending $X_{it}$; $L_{it}$ is our measurement of household labour supply. Note that the changes in relative price of each expenditure category across time are controlled for by the deflating process. Table 3.12 confirms that work-related expenditures are positively correlated with employment status, i.e. household where husband works spends an additional 1.2 percentage points on work-related spending.

Additional robustness checks are used to test if estimation results are altered by excluding spending on food eat away from home. Previous literature documents a drop in food at home expenditure, column (3) and (4) of table 3.13 show no discontinuity on spending both for food at home and food eat away from home, and column (1) gives no drop in total nondurable spending even if excluding food eat away from home[1].

### 3.5.4 Mechanism

#### 3.5.4.1 Family Structure

The literature overwhelmingly shows a drop in food expenditure upon retirement, this paper presents a surprising increase yet not significant in food expenditure. Moreover, food expenditure includes food eat away from home since eating at home and away from home cannot be distinguished in this paper. Table 3.13 has shown that there is no discountinuity in food expenditure even if only look at food expenditure at home. One candidate explanation is a change in the family structure.

In the Chinese tradition, grandchildren are often took care of by their grandparents especially when grandparents are retired. They have more time and energy to care about and cook for their grandchildren while their children are busy with work. Figure 3.8(a) shows no evidence of a significant change in the number of grandchildren upon retirement, but the increasing trend of grandchildren and decreasing trend of children in the household may indicate that incorporating of grandchildren in the households contribute to a relatively stable food consumption.

[1]Information on food eat away from home is only available in year 2002.

#### 3.5.4.2 Elderly Support

In contrast with most western countries that elderly care is offered by states and covered by social security, Chinese culture honours elderly support provided by children especially when children start to work and parents get retired (Banerjee et al., 2010; Giles et al., 2010; Lei et al., 2015). Parents regard their children as a form of saving especially when the credit constraint is tight. It is social norm that retired households expect monetary transfer or in-kind benefits from their children, therefore it may help alleviate a fall in income and smooth consumption.

China Household Income Project (CHIP) 2007 shows that 53.40% of people residing in urban areas, age between 50 and 70, have no pension. China Health and Retirement Longitudinal Study (CHARLS) 2011 shows that 45.17% of household, with household head age between 50 and 70, are receiving monetary transfer from their non-coresident children, and 68.27% of people report that children is the primary instrument for old-age support and 21.66% will rely on pension. Figure 3.8(b) shows a significant increase of family transfer from children onset of retirement.

#### 3.5.4.3 Intra-household Time Allocation

A complementary indicator for testing retirement consumption puzzle is the time use. Aguiar and Hurst (2005) highlight the difference between consumption and expenditure, as stated by Becker (1965), consumption is the output of market expenditure and time. Assuming individual's opportunity cost of time declines upon retirement, individuals will substitute market expenditure with time used for food production. For the purpose of understanding the real change in food consumption, this study provides suggestive evidence on food shopping time based on the time use data from the China Health and Nutrition Survey (CHNS).

The CHNS is an ongoing open cohort, which covers a span of 22 years in 15 provinces and municipal cities that varies in economic development and geography. The survey is consists of about 7200 households with over 30000 individuals drawing from a multistage, random cluster process. The survey provides information on food shopping and food preparation frequency and time use. For example, it asks two types of questions related to food shopping, one type is whether the individual did food shopping in the past week, the other type is the amount of time spent per day on food shopping. The sample is restricted to urban households with male household heads, age between 50 and 70.

To provide a comparison of the CHNS and CHIP data, figure 3.9 presents a jump of the fraction of retirees at age cutoff 60 of 30 percent. This is slightly lower than the CHIP sample, probably due to different sample coverage. Furthermore, balance tests[1] detect no discontinuities in covariates.

Results are provided for the time use of household head and all household members, respectively. Table 3.14 shows no clear change in food production time upon retirement both at household head level and household level. Table 3.9 suggests the existence of heterogeneity in retirement effects by household wealth, proxied by education level. Further estimations are performed to test the heterogeneous pattern in food production pattern. Figure 3.10 presents a jump in individual food shopping time for the low educated but no jump for the high educated household heads. However, there is no clear change in food shopping time at household level for both groups of households. Same applies to food preparing time as it can be observed from figure 3.11. Thus retirees adjust their individual time use to maintain a stable household food shopping time. This provides evidence that households not only able to smooth food expenditure but also food consumption.

#### 3.5.4.4 High household saving rate

Accompany with China's high economic growth rate, the national saving rate is persistently high and accounts for 34% to 53% of the GDP in the past three decades, higher than other Asian countries with similar GDP per capita. In addition, household saving experiences the highest growth among the three sectors(government saving, corporate saving and household saving) since the economy reform in 1978.

Chamon and Prasad (2010) shows a 7 percentage points rise in Chinese urban household savings rate from 1995 to 2005, accounting for one-quarter of disposable income. While the saving rates rise across all demographic groups, the age profile of saving gradually turns into a U-shaped pattern with the younger and the older households saving relatively more. The households in my sample are those age from 50-70, who are in their 20s and 30s as China transiting into a market economy where they bore the most of the increase in uncertainty during the transition but also benefited from the rapid income growth. The mean saving rate is 37 percent in my sample, if the high saving rates prepare households to future adverse shocks, one would expect that households with higher saving rates to bear less of the retirement shock. Surprisingly, table 3.15 shows no significant differences in the total expenditure, food and core

[1]Table of balance test based on CHNS data are provided on request.

nondurable expenditure between high and low saving groups. Given the saving rates of elderly households are high in general, even the relatively lower saving rate group are able to smooth out the shock.

#### 3.5.4.5 Unexpected Retirement

The life cycle hypothesis implies that consumption should not fall with expected retirement, therefore a drop in consumption at an unexpected retirement would not contradict the LCH. However, most of the literature provides evidence on retirement consumption puzzle without distinguishing between expected and unexpected retirement, it is very likely that the conclusion drawn from those analyses are biased upwards. Haider and Stephens (2007) show that consumption drop estimated from expected retirement are roughly a third less than those from other instrument variables such as age.

This study adopts an instrument variable method as a complementary approach to confirm the non-existence of the retirement consumption puzzle in China. In specific, I use subjective retirement expectations as an instrument for actual retirement (Haider and Stephens, 2007). Subjective expectation has been proved to be a strong predictor of the subsequent outcomes (Dominitz, 1998; Stephens Jr, 2004). Implicitly I assume that household expectation error is not correlated with any past information possessed by the household. The nature of this instrument provides potential mechanism of retirement consumption smoothing of elderly households. Hence, this section tests whether consumption falls at expected retirement.

In order to exploit the information on retirement expectation, I use the China Health and Retirement Longitudinal Study (CHARLS). This dataset collects a nationally representative sample of Chinese ages 45 and older, covering a wide range of information on demographic background, health status, household income, wealth and consumption. It resembles the American Health and Retirement Study(HRS) and the English Longitudinal Study of Ageing(ELSA) in terms of sampling and questionnaire. Most importantly, it provides information on expected retirement age "At what age do you plan to stop working? Stopping work in this context shall refer to having stopped all income-related activities and unpaid family business".

Table 3.16 shows a strong first stage effect that an expectation to retire next year increase the actual retirement at next year by 10 percent. The 2SLS estimator shows no evidence of a fall in any subcomponent of consumption at an expected retirement. While it is consistent with the RD finding with respect to total nondurable spending, it contradicts the results on sub-aggregated categories. This could be due to the fact that the RD setting captures the effects on unexpected retirement in addition to the expected effects estimated from the CHARLS data. If the change in expenditure is caused by unexpected event, then the evidence does not refute the life cycle hypotheses. Another possibility is that the repeated cross-sectional data fails to control for some unobserved family traits that would bias the results.

## 3.6 Conclusion

This study firstly documents that there are substantial heterogeneities across subcomponents of nondurable goods in the life cycle pattern of mean and variance of expenditures in the context of China. Specifically, food expenditure presents a hump-shaped profile but the decline after around age 60 is minor, in other words, food expenditure at later stage of life cycle remains relatively stable. Core nondurable expenditure shows a typical hump-shaped life cycle profile, while work-related expenditure displays a significant drop at age 45 and older. This provides insight into the fact that the fall of total nondurable expenditure at age 55 is most likely to be driven by the change in the work-related spending.

Based on the empirical pattern observed above, I tested how much of the adjustments of consumption across time and categories can be attributed to the labour supply status. In particular, a regression discontinuity approach confirms the non-existence of the retirement consumption puzzle. There is a significant positive effect of labour supply on work-related expenditures, but no evidence of a retirement consumption puzzle. In particular, this analysis provide new evidence that elderly households are able to smooth food consumption upon retirement by testing food production theory using time use data from CHNS.

This paper differs from previous literature by exploring several potential explanations for the differences in the profiles across expenditure categories. Due to potential change in family composition and stable household food production time, food consumption smooths upon retirement; work related expenditure drops upon retirement due to the complementary effects of labour supply; total nondurable expenditures stay stable leaving out sub-expenditures that are inputs into market work and amenable to home production. There is no evidence of heterogeneity between households with different education, work units and saving rates. One exception comes from that households of higher education or high saving rates show larger discontinuity in work related expenditure, but it is not inconsistent with the life cycle theory. Additionally, high saving rates of Chinese elderly households provide a possible consumption smoothing mechanism.

Furthermore, by using subjective retirement expectation as an instrumental variable for actual retirement, I find no evidence of a fall in any subcomponent of consumption upon an expected retirement. It implies that the adjustments in sub-aggregated expenditure categories are resulting from unexpected retirement. This, again, is consistent with the LCH. It is worth testing further to what extent that the unexpected retirement is due to health conditions.

This paper also provides an implicit test of the consumption theory that the life cycle profile of consumption goods that are substitution to working status should be different from those goods that are complementary or amenable to home production. Results confirm the complementary mechanism, and the home production theory finds support in adjustments of intra-household food shopping time.

As with any study, there are some limitations of this paper. First, this study use repeated cross section data that may fail to control for some unobserved household characteristics that would bias the results. Moreover, while the data covers a relatively long time span ranging from 1995 to 2007, there is only four waves of cross section data, the analysis is likely to suffer from some transitory shocks that is difficult to separate from time trends.

Nonetheless, the estimates and their magnitude are important for understanding the life cycle profile of Chinese households' consumption behaviour. Estimation results demonstrate that disaggregated expenditures exhibit substantial heterogeneity with respect to both mean and variance. In particular, work status is closely associated with total nondurable expenditure life cycle profile. The finding has important implications for using disaggregated data to test the existence of retirement consumption puzzle, for testing consumption theories and for providing evidence on the widely debated topic of postponing the mandatory retirement age given the economic deceleration and rapid ageing problems in China.

# Tables

Table 3.1: Summary of Mean Change in Expenditure Over the Life Cycle by Consumption Category

| Disaggregated consumption category | (1) Share of expenditures at age 25-27 | (2) Share of expenditures at age 43-45 | (3) Share of expenditures at age 64-66 | (4) Log change in expenditure between age 25-45 | (5) Log change in expenditure between age45-65 |
|---|---|---|---|---|---|
| Food | 0.56 | 0.60 | 0.64 | 0.28 | 0.00 |
| Work | 0.26 | 0.23 | 0.17 | 0.05 | -0.60 |
| Core | 0.18 | 0.17 | 0.19 | 0.25 | -0.12 |
| Total | 1.00 | 1.00 | 1.00 | 0.21 | -0.14 |

Note: This table summarizes the life cycle mean expenditure profiles for total nondurable good and three main consumption categories: food, work-related and the core nondurables, as shown in figure 3.1(a) and figure 3.1(a). Column 1-3 report the share of expenditures at ages 25-27, 43-45 and 64-66, respectively. I use three year age average to smooth out some of the age-to-age variability. Column 4 and 5 report the log change in mean expenditure between ages 25 and 45, 45 and 65, which is the difference in coefficient on the age dummies from the regression of log expenditure on age dummies and demographic controls (equation 3.1).

Table 3.2: Summary Statistics for the RD Sample

| | Observations | Mean | s.d. |
|---|---|---|---|
| **Demographic Charateristics** | | | |
| age | 5160 | 58.07 | 5.99 |
| retire | 5160 | 0.43 | 0.50 |
| high school or above | 5144 | 0.41 | 0.49 |
| number of household members | 5160 | 3.05 | 1.02 |
| number of children | 5160 | 0.79 | 0.73 |
| couple | 5160 | 0.98 | 0.16 |
| minority | 5160 | 0.04 | 0.18 |
| spouse retired | 5160 | 0.63 | 0.48 |
| housing area | 4032 | 52.97 | 27.80 |
| **Institution Background** | | | |
| share of Public sectors | 5160 | 0.88 | |
| share of Government workers | 324 | 0.11 | |
| share of Instituition workers | 476 | 0.16 | |
| share of State-owned Enterprise workers | 1607 | 0.54 | |
| share of Private Sector Worker | 576 | 0.19 | |
| **Expenditures** | | | |
| food | 5160 | 8575.99 | 7471.67 |
| work-related | 5160 | 3108.58 | 4792.30 |
| core nondurable | 5160 | 3000.91 | 5,560,401 |
| total nondurable | 5160 | 14685.48 | 12930.41 |

Note: This table summarizes demographic characteristics and main expenditure categories for the RD sample, which restricts the sample to households that age between 50 and 70 with age 60 excluded.

The expenditures reported here are measured in Chinese currency Yuan(1 Yuan=0.12 Pound) for the purpose of understanding the scale of spending, while log value of expenditures is used in the regression-based analysis.

Table 3.3: First Stage: The Effect of Mandatory Retirement Age on Retirement

| | (1) Retire | (2) Retire | (3) Retire | (4) Retire | (5) Retire | (6) Retire |
|---|---|---|---|---|---|---|
| D(age>60) | 0.464*** | 0.465*** | 0.350*** | 0.354*** | 0.357*** | 0.360*** |
| | (0.027) | (0.024) | (0.042) | (0.037) | (0.070) | (0.062) |
| age*D | -0.035*** | -0.035*** | -0.051** | -0.052*** | 0.016 | 0.007 |
| | (0.004) | (0.004) | (0.016) | (0.015) | (0.052) | (0.048) |
| $age^2$*D | | | -0.005*** | -0.005*** | -0.005 | -0.005 |
| | | | (0.001) | (0.001) | (0.010) | (0.010) |
| $age^3$*D | | | | | 0.001 | 0.001 |
| | | | | | (0.001) | (0.001) |
| age | 0.040*** | 0.040*** | 0.076*** | 0.075*** | 0.041 | 0.045 |
| | (0.003) | (0.003) | (0.011) | (0.010) | (0.039) | (0.035) |
| $age^2$ | | | 0.003*** | 0.003*** | -0.004 | -0.003 |
| | | | (0.001) | (0.001) | (0.008) | (0.007) |
| $age^3$ | | | | | -0.000 | -0.000 |
| | | | | | (0.000) | (0.000) |
| Province dummies | No | YES | No | YES | No | YES |
| Year dummies | No | YES | No | YES | No | YES |
| Polynomial order | 1st order | 1st order | 2nd order | 2nd order | 3rd order | 3rd order |
| F statistics for joint significance | 199,860 | 251,390 | 124,100 | 148,900 | 31,240 | 39,860 |
| R-sqr | 0.534 | 0.539 | 0.536 | 0.540 | 0.536 | 0.541 |
| N | 5160 | 5160 | 5160 | 5160 | 5160 | 5160 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.
Variable $D$ is an indicator for whether age of household head is above 60; variable $age$ is the difference between real age and age 60; variable $age^2$ and $age^3$ are the square and cubic of $age$, respectively.

Table 3.4: Balance Test: the Effect of Retirement on Predetermined Variables

| | (1) edu_high | (2) hh_no | (3) child_no | (4) couple | (5) minority | (6) retire_sp | (7) hukou |
|---|---|---|---|---|---|---|---|
| Retire | -0.123 | -0.211 | -0.108 | 0.018 | -0.024 | 0.047 | 0.007 |
| | (0.095) | (0.157) | (0.184) | (0.029) | (0.020) | (0.055) | (0.007) |
| age*Retire | | | | | 0.002 | | |
| | | | | | (0.002) | | |
| $age^2$*(1-Retire) | | | | | | -0.004*** | 0.000* |
| | | | | | | (0.000) | (0.000) |
| $age^3$*(1-Retire) | -0.000 | 0.000* | 0.000 | 0.000 | | | |
| | (0.000) | (0.000) | (0.000) | (0.000) | | | |
| age | 0.008 | -0.013 | -0.042** | -0.003 | 0.001 | 0.013* | 0.000 |
| | (0.009) | (0.010) | (0.016) | (0.002) | (0.002) | (0.003) | (0.000) |
| $age^2$ | -0.001 | | 0.001 | -0.000 | | | |
| | (0.001) | | (0.001) | (0.000) | | | |
| Province dummies | YES | YES | YES | YES | YES | YES | YES |
| Year dummies | YES | YES | YES | YES | YES | YES | YES |
| Polynomial order | L3R2 | L3R1 | L3R2 | L3R2 | L1R1 | L2R1 | L2R1 |
| R-sqr | 0.027 | 0.039 | 0.143 | 0.006 | 0.049 | 0.273 | 0.011 |
| N | 5144 | 5160 | 5160 | 5160 | 5160 | 5160 | 5155 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.

All the regressions control for year dummies and province dummies. The polynomial orders are chosen by AIC. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order) for the balance test, for example, L3R2 in column (1) means that third order on the left and second order on the right is chosen by the AIC criteria.

Table 3.5: Second Stage: The Effect of Retirement on Expenditure Categories

| | (1) Total | (2) Food | (3) Work | (4) Core |
|---|---|---|---|---|
| Retire | -0.068 | 0.074 | -0.414** | -0.102 |
| | (0.062) | (0.054) | (0.137) | (0.103) |
| age*Retire | -0.021*** | -0.016*** | | |
| | (0.005) | (0.005) | | |
| $age^2$*(1-Retire) | | | -0.003*** | |
| | | | (0.001) | |
| $age^3$*(1-Retire) | | | | 0.000*** |
| | | | | (0.000) |
| age | 0.009* | 0.006 | -0.025** | -0.004 |
| | (0.005) | (0.004) | (0.009) | (0.007) |
| Province dummies | YES | YES | YES | YES |
| Year dummies | YES | YES | YES | YES |
| Polynomial order | L1R1 | L1R1 | L2R1 | L3R1 |
| R-sqr | 0.416 | 0.424 | 0.298 | 0.248 |
| N | 5160 | 5160 | 5160 | 5160 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.
All the regressions control for year dummies and province dummies. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order), for example, L3R2 in column (1) means that third order on the left and second order on the right is chosen by the AIC criteria.

Table 3.6: Robustness Check: the Effect of Retirement based on Different Age Band

| | (1) Total | (2) Food | (3) Work | (4) Core |
|---|---|---|---|---|
| 45-75 | -0.136 | 0.023 | -0.479*** | -0.202 |
| | (0.076) | (0.073) | (0.105) | (0.114) |
| N | 7863 | 7863 | 7863 | 7863 |
| 46-74 | -0.102 | 0.029 | -0.477*** | -0.172 |
| | (0.082) | (0.081) | (0.107) | (0.125) |
| N | 7308 | 7308 | 7308 | 7308 |
| 47-73 | -0.089 | 0.050 | -0.473*** | -0.148 |
| | (0.090) | (0.087) | (0.116) | (0.294) |
| N | 6828 | 6828 | 6828 | 6828 |
| 48-72 | -0.045 | 0.094 | -0.451*** | -0.151 |
| | (0.096) | (0.091) | (0.121) | (0.173) |
| N | 6273 | 6273 | 6273 | 6273 |
| 49-71 | -0.025 | 0.065 | -0.377** | -0.195 |
| | (0.100) | (0.089) | (0.130) | (0.154) |
| N | 5691 | 5691 | 5691 | 5691 |
| 51-69 | 0.095 | 0.069 | -0.337 | -0.089 |
| | (0.109) | (0.061) | (0.217) | (0.106) |
| N | 4586 | 4586 | 4586 | 4586 |
| 52-68 | -0.041 | 0.085 | -0.356* | -0.121 |
| | (0.073) | (0.067) | (0.159) | (0.113) |
| N | 4105 | 4105 | 4105 | 4105 |
| 53-67 | 0.000 | 0.137 | -0.210 | -0.101 |
| | (0.084) | (0.074) | (0.246) | (0.134) |
| N | 3557 | 3557 | 3557 | 3557 |
| 54-66 | -0.006 | 0.098 | -0.332 | -0.130 |
| | (0.196) | (0.169) | (0.187) | (0.203) |
| N | 2990 | 2990 | 2990 | 2990 |
| 55-65 | 0.113 | 0.197 | -0.314 | -0.043 |
| | (0.200) | (0.167) | (0.218) | (0.255) |
| N | 2526 | 2526 | 2526 | 2526 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %. All the regressions control for year dummies and province dummies. The polynomial orders are chosen by AIC.

Table 3.7: Robustness Check: the Effect of Retirement on Expenditure Categories—Quadratic Specification

| | (1) Total | (2) Food | (3) Work | (4) Core |
|---|---|---|---|---|
| Retire | 0.107 | 0.200 | -0.374 | -0.001 |
| | (0.133) | (0.118) | (0.272) | (0.206) |
| age*Retire | 0.010 | 0.007 | 0.064 | 0.002 |
| | (0.025) | (0.024) | (0.053) | (0.039) |
| $age^2$*Retire | 0.005* | 0.003 | 0.006 | 0.005 |
| | (0.002) | (0.002) | (0.005) | (0.004) |
| age | -0.032 | -0.024 | -0.068 | -0.020 |
| | (0.019) | (0.020) | (0.038) | (0.030) |
| $age^2$ | -0.003* | -0.002 | -0.007* | -0.003 |
| | (0.001) | (0.002) | (0.003) | (0.002) |
| Province dummies | YES | YES | YES | YES |
| Year dummies | YES | YES | YES | YES |
| R-sqr | 0.386 | 0.396 | 0.290 | 0.235 |
| N | 5160 | 5160 | 5160 | 5160 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %. All the regressions control for year dummies and province dummies.

Table 3.8: Robustness Check: Placebo Test

| | Left-side subsample "fake" cutoff—54 | | | | Right-side subsample "fake" cutoff—64 | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | Total | Food | Work | Core | Total | Food | Work | Core |
| Retire | -0.060 | 0.509 | -0.208 | -0.737 | -0.061 | 0.149 | -0.526 | -0.516 |
| | (0.741) | (0.743) | (1.6550 | (1.013) | (0.280) | (0.256) | (0.603) | (0.526) |
| age*Retire | | | 0.087 | | | | | |
| | | | (0.559) | | | | | |
| $age^2$*Retire | | | -0.041 | | | | | |
| | | | (0.067) | | | | | |
| $age^3$*(1-Retire) | -0.001 | -0.001 | | 0.001 | 0.004 | 0.001 | 0.005 | 0.011 |
| | (0.001) | (0.001) | | (0.003) | (0.004) | (0.004) | (0.010) | (0.008) |
| age | 0.022 | -0.001 | 0.013 | 0.035 | -0.017 | -0.012 | -0.035 | -0.023 |
| | (0.036) | (0.035) | (0.026) | (0.069) | (0.009) | (0.009) | (0.021) | (0.015) |
| $age^2$ | -0.003 | -0.003 | | | | | | |
| | (0.002) | (0.002) | | | | | | |
| R-sqr | 0.408 | 0.310 | 0.257 | 0.165 | 0.411 | 0.431 | 0.248 | 0.192 |
| N | 3272 | 3272 | 3272 | 3272 | 2146 | 2146 | 2146 | 2146 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.
All the regressions control for year dummies and province dummies.
The polynomial orders are chosen by AIC. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order), for example, *$age^2$*Retire* means that second order polynomial specification is chosen by the AIC criteria for the right-side approximation, *$age^2$*(1-Retire)* means that second order polynomial specification is chosen for the left-side approximation.

Table 3.9: Robustness Check: the Effect of Retirement on Different Education Group

| | Non-high school group | | | | High_school group | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | Total | Food | Work | Core | Total | Food | Work | Core |
| Retire | -0.010 | -0.030 | -0.349 | -0.014 | -0.038 | 0.106 | -0.375* | -0.156 |
| | (0.086) | (0.160) | (0.218) | (0.147) | (0.133) | (0.122) | (0.147) | (0.130) |
| age*Retire | | -0.012 | | | | | | |
| | | (0.016) | | | | | | |
| $age^2$*Retire | | -0.001 | | | | | | |
| | | (0.002) | | | | | | |
| $age^2$*(1-Retire) | -0.002*** | | | | | | -0.002* | |
| | (0.000) | | | | | | (0.001) | |
| $age^3$*(1-Retire) | | | 0.000*** | 0.000*** | 0.000 | 0.000 | | 0.000 |
| | | | (0.000) | (0.000) | (0.000) | (0.000) | | (0.000) |
| age | -0.013* | 0.012 | -0.034* | -0.011 | -0.001 | -0.002 | -0.015 | 0.004 |
| | (0.005) | (0.009) | (0.014) | (0.009) | (0.012) | (0.012) | (0.011) | (0.010) |
| $age^2$ | | | | | -0.000 | -0.000 | | |
| | | | | | (0.001) | (0.001) | | |
| R-sqr | 0.413 | 0.412 | 0.292 | 0.263 | 0.421 | 0.454 | 0.306 | 0.231 |
| N | 3016 | 3016 | 3016 | 3016 | 2128 | 2128 | 2128 | 2128 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.
All the regressions control for year dummies and province dummies.
The polynomial orders are chosen by AIC. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order), for example, *$age^2$ *Retire* means that second order polynomial specification is chosen by the AIC criteria for the right-side approximation, *$age^2$ *(1-Retire)* means that second order polynomial specification is chosen for the left-side approximation.

Table 3.10: Robustness Check: the Effect of Retirement on Different Work unit[a]

| | Public sector | | | | Private sector | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | Total | Food | Work | Core | Total | Food | Work | Core |
| Retire | -0.026 | 0.121 | -0.459** | -0.045 | -3,833 | -2,584 | -5,476 | -2,216 |
| | (0.123) | (0.121) | (0.167) | (0.194) | (15,988) | (12,444) | (74,525) | (5,223) |
| age*Retire | | | | | | | 3,574 | 0.373 |
| | | | | | | | (5,812) | (2,018) |
| $age^2$ *Retire | | | | | | | 0.128 | 0.072 |
| | | | | | | | (1,490) | (0.076) |
| $age^3$ *Retire | | | | | | | -0.038 | -0.009 |
| | | | | | | | (0.119) | (0.018) |
| $age^2$ *(1-Retire) | | | -0.004*** | | | | | |
| | | | (0.001) | | | | | |
| $age^3$ *(1-Retire) | 0.000 | 0.000 | | 0.000 | -0.000 | -0.000 | | |
| | (0.000) | (0.000) | | (0.000) | (0.002) | (0.001) | | |
| age | -0.009 | -0.010 | -0.029* | -0.010 | 0.049 | 0.043 | -0.280 | -0.031 |
| | (0.010) | (0.011) | (0.011) | (0.017) | (0.259) | (0.202) | (0.566) | (0.156) |
| $age^2$ | -0.001 | -0.000 | | -0.000 | | | -0.030 | 0.003 |
| | (0.001) | (0.001) | | (0.001) | | | (0.060) | (0.029) |
| $age^3$ | | | | | | | | 0.001 |
| | | | | | | | | (0.001) |
| R-sqr | 0.381 | 0.353 | 0.297 | 0.273 | . | 0.265 | . | . |
| N | 3962 | 3962 | 3962 | 3962 | 505 | 505 | 505 | 505 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.
All the regressions control for year dummies and province dummies.
The polynomial orders are chosen by AIC. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order), for example, *$age^2$ *Retire* means that second order polynomial specification is chosen by the AIC criteria for the right-side approximation, *$age^2$ *(1-Retire)* means that second order polynomial specification is chosen for the left-side approximation.
[a] This paper defines two broad categories of workunit: public and private sectors. Public sector is a broad term, for the purpose of distinguishing from private sector. Public sector includes government organizations, institutions, state-owned and collective-owned enterprises. Private sector includes private firms, sino-foreign joint ventures, foreign companies, state share-holding companies, and other ownership.

Table 3.11: Robustness Check: the Effect of Retirement across time

| | (1) Total | (2) Food | (3) Work | (4) Core |
|---|---|---|---|---|
| Year==1995 | -0.029 | -0.022 | -0.603* | -0.127 |
| | (0.160) | (0.141) | (0.316) | (0.266) |
| N | 1557 | 1557 | 1557 | 1557 |
| Year==1999 | 0.041 | 0.069 | -0.285 | -0.125 |
| | (0.244) | (0.118) | (0.426) | (0.421) |
| N | 887 | 887 | 887 | 887 |
| Year==2002 | -0.067 | -0.015 | -0.160 | 0.079 |
| | (0.180) | (0.155) | (0.245) | (0.158) |
| N | 1588 | 1588 | 1588 | 1588 |
| Year==2007 | -0.109 | 0.066 | -0.214 | -0.319 |
| | (0.157) | (0.267) | (0.274) | (0.433) |
| N | 1128 | 1128 | 1128 | 1128 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %. All the regressions control for year dummies and province dummies. The polynomial orders are chosen by AIC.

Table 3.12: Robustness Check: the Relationship Between Work status and Spending by Consumption Category

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Share of food | Share of food | Share of work-related | Share of work-related | Share of core-non | Share of core-non |
| employ | -0.018** | -0.015* | 0.019*** | 0.012** | -0.001 | 0.002 |
| | (0.006) | (0.006) | (0.003) | (0.003) | (0.005) | (0.005) |
| Model specification | OLS | 2SLS | OLS | 2SLS | OLS | 2SLS |
| R-sqr | 0.228 | 0.221 | 0.240 | 0.177 | 0.090 | 0.080 |
| N | 5418 | 5377 | 5418 | 5377 | 5418 | 5377 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %. This table provides estimation from demand system analysis, where total expenditure is instrumented with income and education.

Table 3.13: Robustness Check: the Effect of Retirement on Alternative Measurement of Food Expenditures

| | (1) Total | (2) Food | (3) Food away | (4) Food home |
|---|---|---|---|---|
| retire | -0.067 | -0.015 | 0.018 | -0.027 |
| | (0.180) | (0.155) | (0.553) | (0.138) |
| $age^3$*(1-Retire) | -0.000 | -0.000 | -0.000 | -0.000 |
| | (0.000) | (0.000) | (0.001) | (0.000) |
| age | 0.004 | 0.011 | -0.026 | 0.015 |
| | (0.016) | (0.014) | (0.051) | (0.012) |
| $age^2$ | -0.002 | -0.002* | -0.001 | -0.002** |
| | (0.001) | (0.001) | (0.004) | (0.001) |
| R-sqr | 0.199 | 0.274 | 0.161 | 0.254 |
| N | 1588 | 1588 | 1484 | 1588 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.

Information on food eat away from home is only available in year 2002.

All the regressions control for year dummies and province dummies.

The polynomial orders are chosen by AIC. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order), for example, *$age^2$ *Retire* means that second order polynomial specification is chosen by the AIC criteria for the right-side approximation, *$age^2$ *(1-Retire)* means that second order polynomial specification is chosen for the left-side approximation.

Table 3.14: Mechanism: the Effect of Retirement on Food Production time

| | (1) Household head time spend on buying food | (2) Household head time spend on preparing food | (3) Household time spend on buying food | (4) Household time spend on preparing food |
|---|---|---|---|---|
| Retire | 13.097 | -1.968 | -24.542 | 1.375 |
| | (12.853) | (14.055) | (19.870) | (11.370) |
| age*Retire | | -0.182 | | -0.089 |
| | | (2.267) | | (0.684) |
| $age^2$*Retire | | 0.150 | | |
| | | (0.256) | | |
| $age^3$*(1-Retire) | -0.010 | | 0.010 | |
| | (0.007) | | (0.007) | |
| age | 0.529 | 0.205 | 1.119 | 0.189 |
| | (0.645) | (0.491) | (0.947) | (0.621) |
| Province dummies | YES | YES | YES | YES |
| Year dummies | YES | YES | YES | YES |
| N | 1657 | 1566 | 3425 | 3425 |
| R-sq | 0.240 | 0.573 | 0.129 | 0.622 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %. All the regressions control for year dummies and province dummies.
The time use data are based on the CHNS.
The polynomial orders are chosen by AIC. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order), for example, $age^2$ **Retire* means that second order polynomial specification is chosen by the AIC criteria for the right-side approximation, $age^2$ **(1-Retire)* means that second order polynomial specification is chosen for the left-side approximation.The polynomial orders are chosen by AIC.

Table 3.15: Mechanism: the Effect of Retirement on Different Saving Rate Group

| | Low saving rate group | | | | High saving rate group | | | |
|---|---|---|---|---|---|---|---|---|
| | Total | Food | Work | Core | Total | Food | Work | Core |
| retire | -0.072 | 0.008 | -0.299 | -0.221 | -0.084 | -0.024 | -0.462** | -0.123 |
| | (0.178) | (0.163) | (0.202) | (0.154) | (0.070) | (0.243) | (0.164) | (0.130) |
| age*Retire | | | | | -0.026*** | 0.000 | | |
| | | | | | (0.007) | (0.067) | | |
| $age^2$*Retire | | | | | | -0.005 | | |
| | | | | | | (0.014) | | |
| $age^3$*Retire | | | | | | 0.000 | | |
| | | | | | | (0.001) | | |
| $age^2$*(1-Retire) | | | | -0.002 | | | -0.004*** | -0.002** |
| | | | | (0.001) | | | (0.001) | (0.001) |
| $age^3$*(1-Retire) | 0.000 | -0.000 | 0.000*** | | | | | |
| | (0.000) | (0.000) | (0.000) | | | | | |
| age | -0.006 | 0.001 | -0.033* | 0.003 | 0.010 | 0.008 | -0.020 | -0.005 |
| | (0.014) | (0.013) | (0.013) | (0.009) | (0.005) | (0.007) | (0.011) | (0.009) |
| $age^2$ | -0.000 | -0.001 | | | | | | |
| | (0.001) | (0.001) | | | | | | |
| R-sqr | 0.420 | 0.491 | 0.306 | 0.238 | 0.444 | 0.404 | 0.298 | 0.275 |
| N | 2569 | 2569 | 2569 | 2569 | 2566 | 2566 | 2566 | 2566 |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.

All the regressions control for year dummies and province dummies.

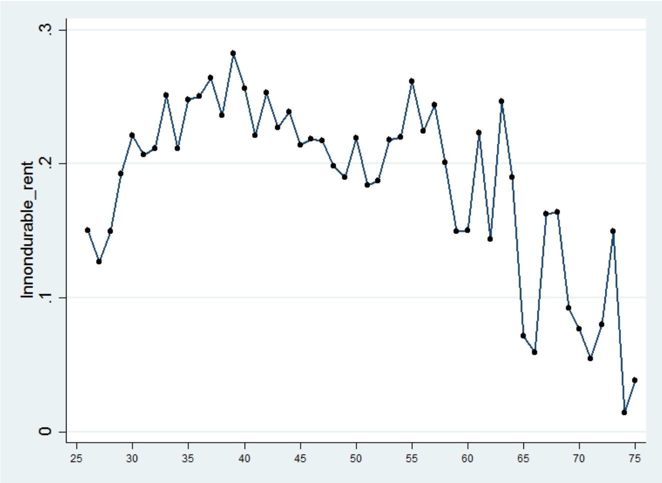
The polynomial orders are chosen by AIC. The AIC criteria chooses among symmetric and asymmetric polynomial order(up to the fifth order), for example, *$age^2$*Retire* means that second order polynomial specification is chosen by the AIC criteria for the right-side approximation, *$age^2$*(1-Retire)* means that second order polynomial specification is chosen for the left-side approximation.The polynomial orders are chosen by AIC.

Table 3.16: Mechanism: the Effect of Expected Retirement on Expenditure Categories

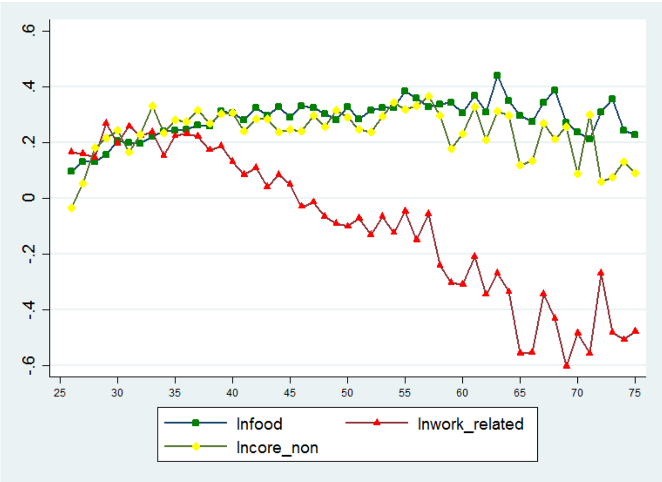
| | (1)<br>change in<br>food<br>expenditure | (2)<br>change in<br>work-related<br>expenditure | (3)<br>change in<br>core-non<br>expenditure | (4)<br>change in<br>total<br>expendituree |
|---|---|---|---|---|
| **OLS** | | | | |
| actual retirement | -0.048 | -0.097* | -0.024 | -0.107** |
| at year t+1 | (0.050) | (0.058) | (0.050) | (0.047) |
| N | 6045 | 5647 | 6100 | 5091 |
| R-sq | 0.018 | 0.044 | 0.036 | 0.037 |
| **2SLS** | | | | |
| actual retirement | 0.244 | -0.798 | 1.649 | 0.316 |
| at year t+1 | (0.980) | (0.906) | (1.088) | (0.795) |
| N | 4153 | 3956 | 4243 | 3585 |
| | actual retirement at year t+1 | | | |
| **First Stage** | | | | |
| expected year t+1 | 0.102*** | | | |
| retirement at year t | (0.022) | | | |
| F-test | 20.52 | | | |

Note: Standard errors are clustered at province-age level. $***$ significant at 1 %; $**$ significant at 5 %; $*$ significant at 10 %.
This analysis is based on the China Health and Retirement Longitudinal Study (CHARLS). Subjective retirement expectations are used as an instrument for actual retirement. Each column shows the effect of retirement on the change of main subaggregated categories and total nondurable expenditure, respectively.

# Figures



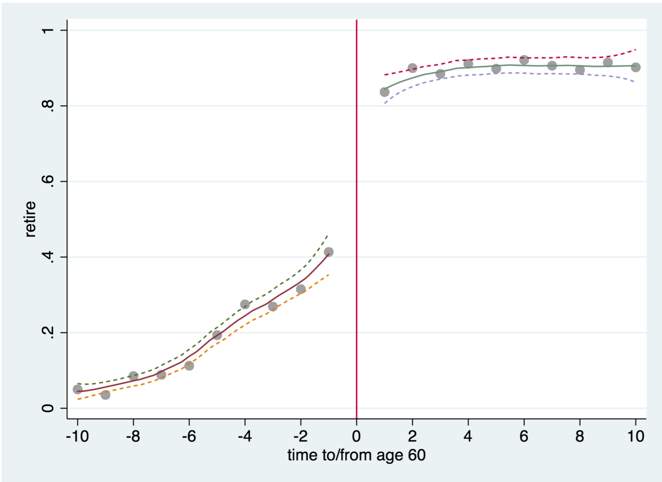

(a) Mean Profile



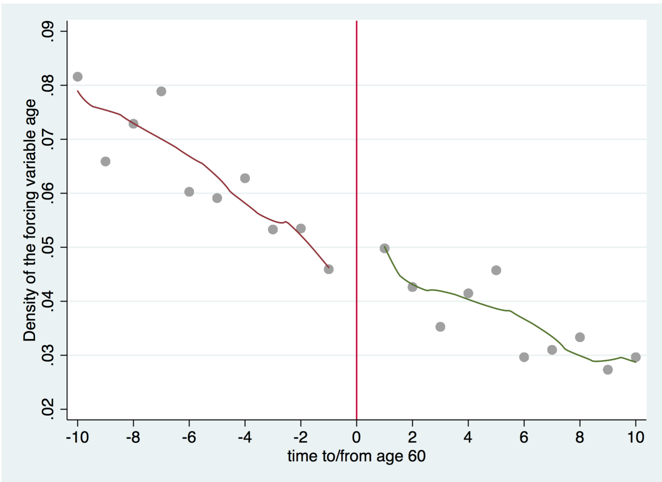

(b) Variance Profile

Figure 3.1: Life Cycle Pattern of Total Non-durable Expenditure

Note: The regression based nondurable expenditures profile have been shown in graph (a) and (b), respectively. The age effects are given by controlling for year and family composition controls, with age 25 as the baseline group. Due to few extreme values at age 26, the coefficient plot at age 26 exhibits a big variation. Graph (a) gives the evolution of mean of total nondurable expenditure across life cycle, while total nondurable expenditure is disaggregated to three main categories in graph(b).



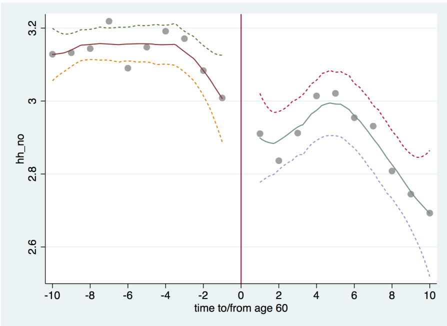

(a) Fraction of retirees

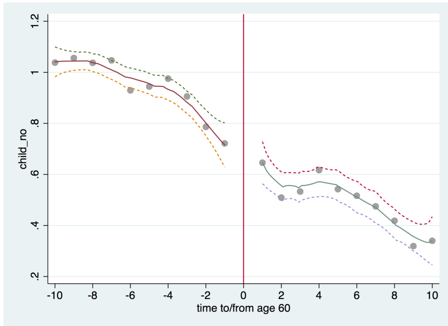



(b) Density of the running variable

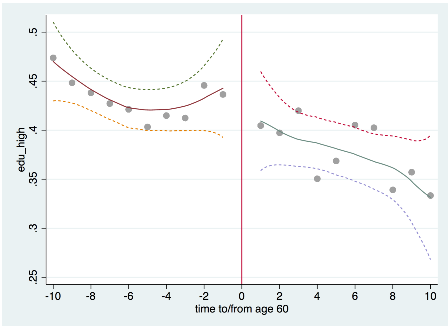
Figure 3.2: First Stage—The Effect of Mandatory Retirement Age on Retirement

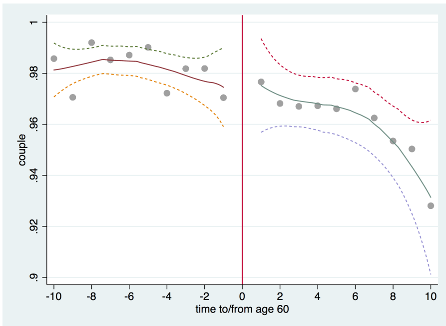
Note: Panel(a): Cells are the proportion of people in each age from age 50 to 70. The solid lines and dashed lines are local linear fitted outcomes and associated confidence intervals, respectively, on either side of the age 60 cutoff.
Panel(b): It plots the density of households at each age over the total number of households with household age between 50 and 70.



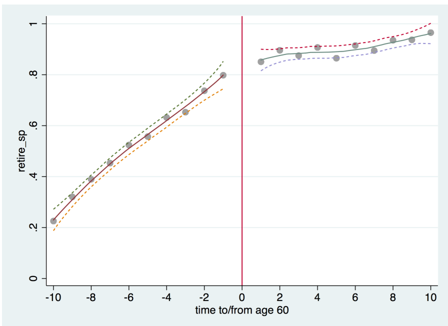


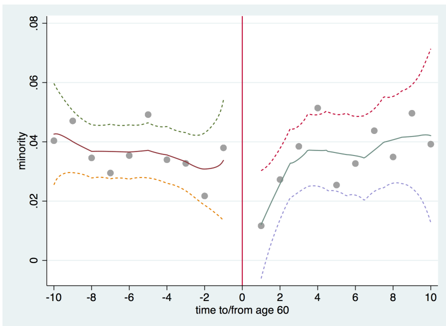

(a) Household Size (b) Number of Children






(c) Above high school education (d) couple

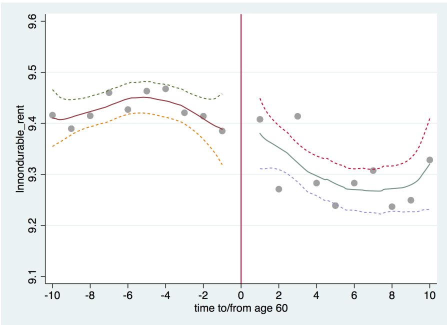




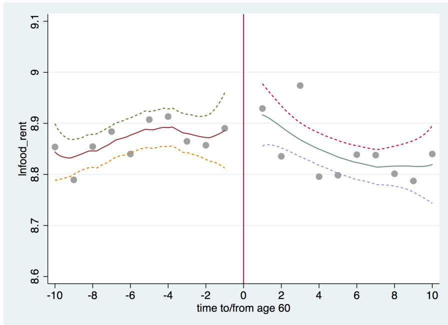

(e) Retirement Status of the Spouse (f) Minority

Figure 3.3: Balance Test—The Effect of Retirement on Predetermined Variables

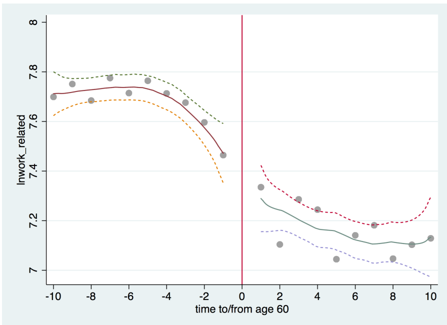
Note: Cells are the household head in each age from age 50 to 70. The solid lines and dashed lines are local linear fitted outcomes and associated confidence intervals, respectively, on either side of the age 60 cutoff.



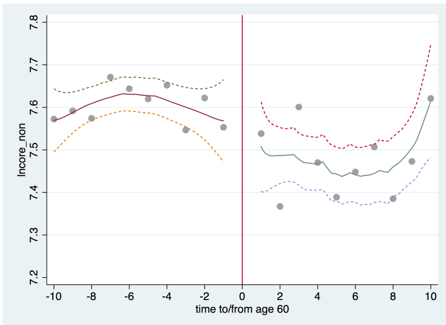

(a) Total nondurable expenditures




(b) food expenditures




(c) work related expenditures




(d) Core-nondurable expenditures

Figure 3.4: Reduced Form—The Effect of Retirement on Main Expenditure Categories

Note: Cells are the household head in each age from age 50 to 70. The solid lines and dashed lines are local linear fitted outcomes and associated confidence intervals, respectively, on either side of the age 60 cutoff.

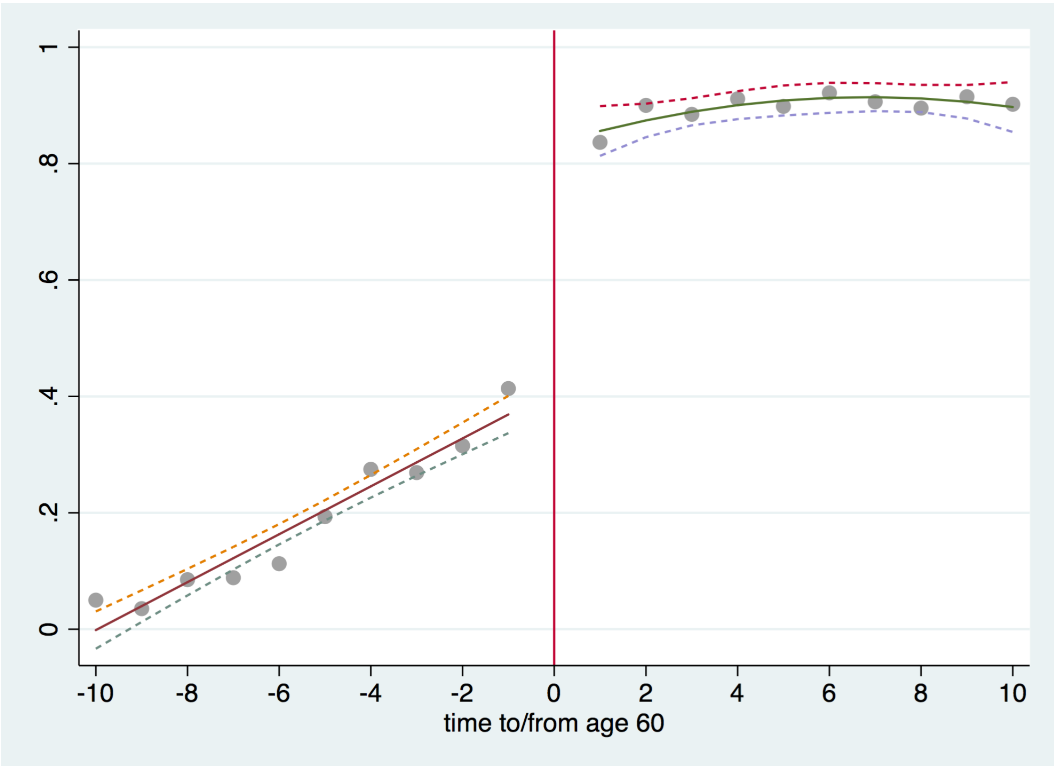
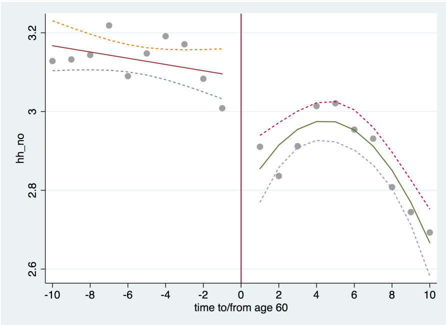
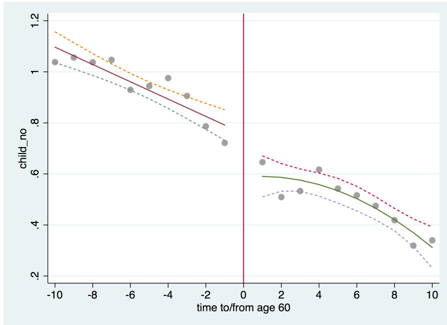



Figure 3.5: Robustness Check: Quadratic Polynomial Regressions — the First Stage

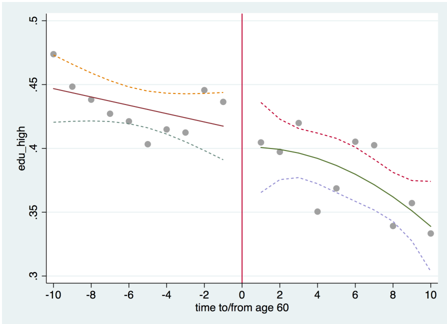
Note: Cells are the proportion of people in each age from age 50 to 70. The solid lines and dashed lines are the predicted outcomes and associated confidence intervals, respectively.



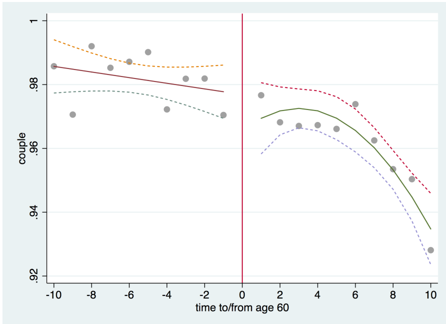

(a) Household Size



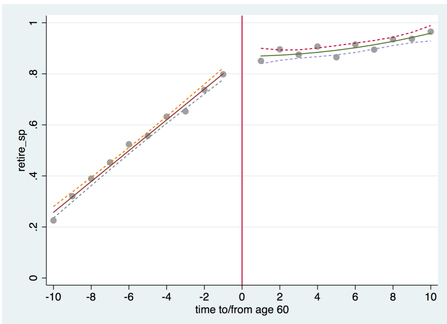

(b) Number of Children



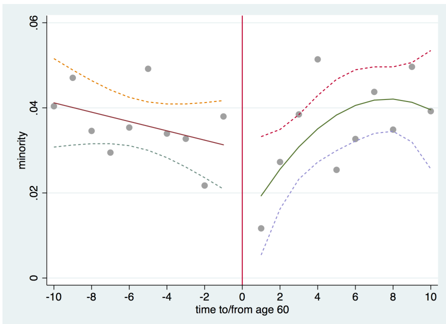

(c) Above high school education




(d) couple



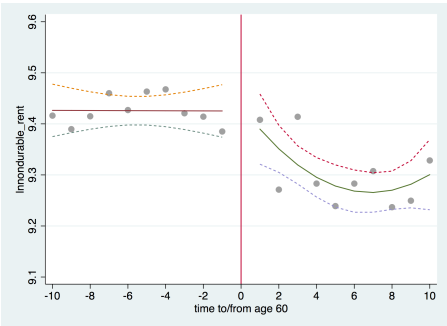

(e) Retirement Status of the Spouse



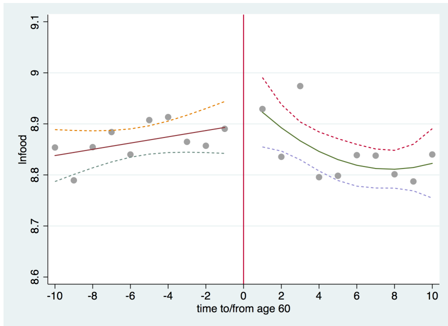

(f) Minority

Figure 3.6: Robustness Check: Quadratic Polynomial Regressions — Balance Test

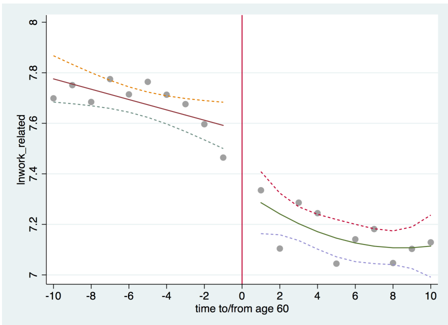
Note: Cells are the household head in each age from age 50 to 70. The solid lines and dashed lines are the predicted outcomes and associated confidence intervals, respectively.



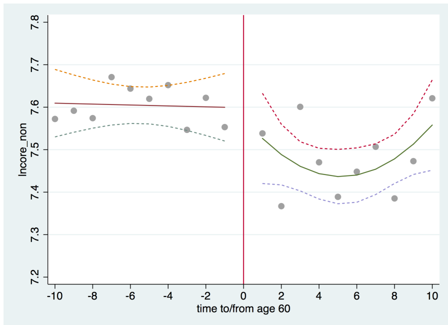

(a) Total nondurable expenditures




(b) food expenditures



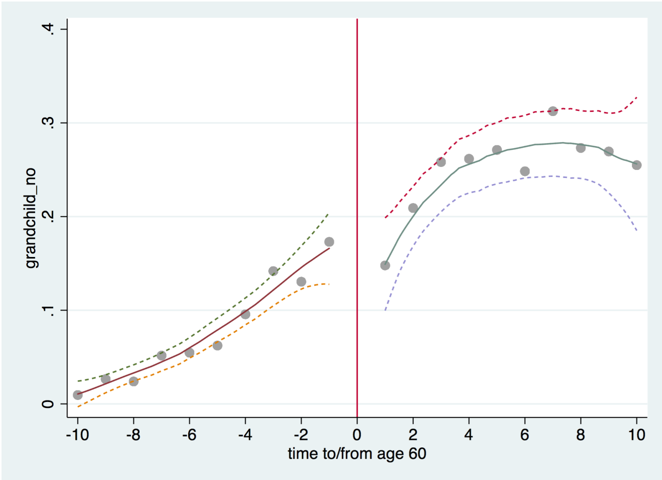

(c) work related expenditures



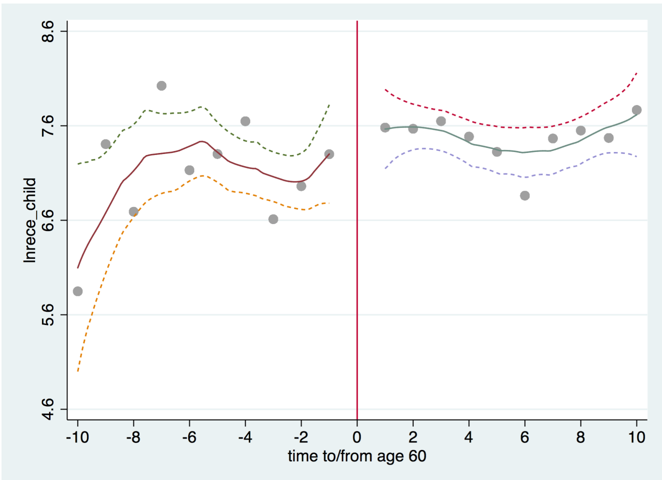

(d) Core-nondurable expenditures

Figure 3.7: Robustness Check: Quadratic Polynomial Regressions — Reduced Form Estimation

Note: Cells are the household head in each age from age 50 to 70. The solid lines and dashed lines are the predicted outcomes and associated confidence intervals, respectively.




(a) the number of grandchildren living with household head




(b) the amount of family transfer from non-coresident children

Figure 3.8: Mechanism: Household Composition and Family Transfer

Note: Panel (a) shows local linear estimation of the number of grandchildren living with household heads. Panel (b) shows local linear estimation of the logged amount of family transfer from non-coresident children.

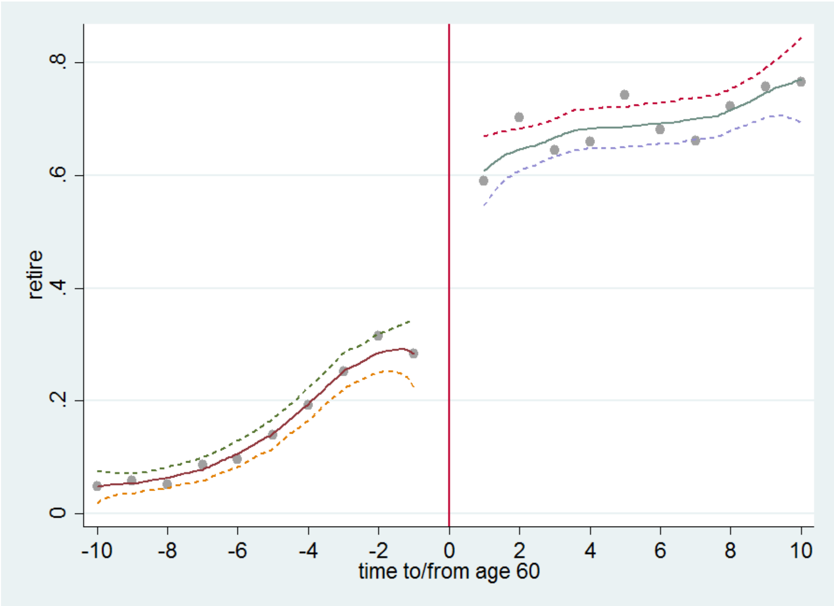
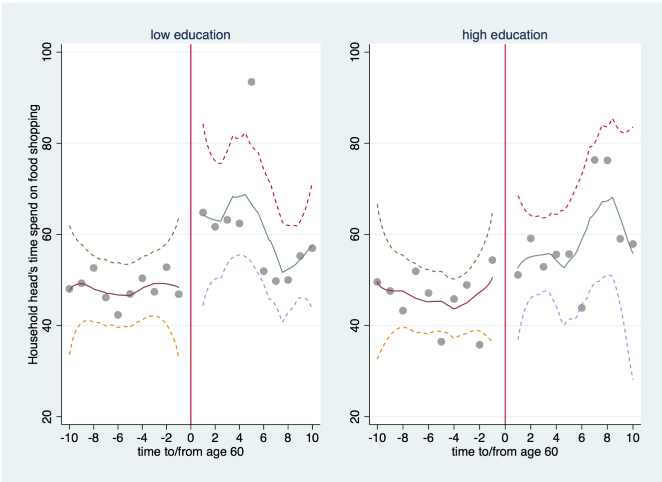
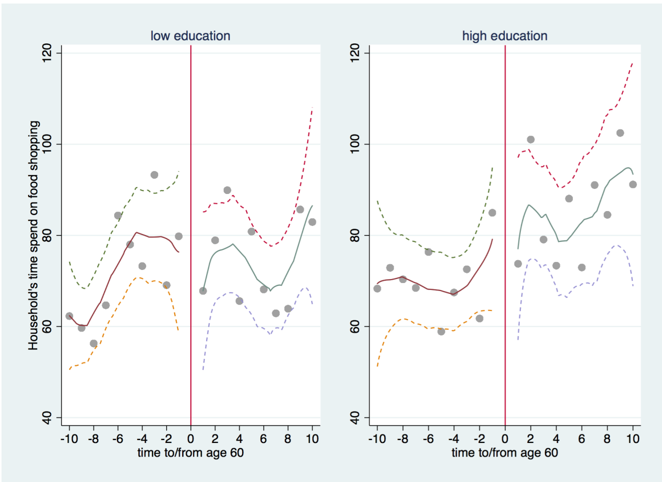



Figure 3.9: Mechanism: The Fraction of Retirees in the CHNS Sample

Note: It shows local linear estimation of fraction of retirees in the CHNS sample. The solid lines and dashed lines are local linear fitted outcomes and associated confidence intervals, respectively, on either side of the age 60 cutoff.



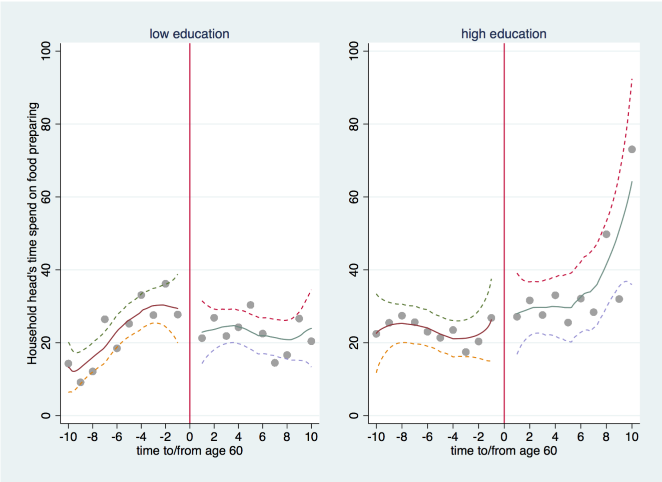

(a) Food shopping time of the individual



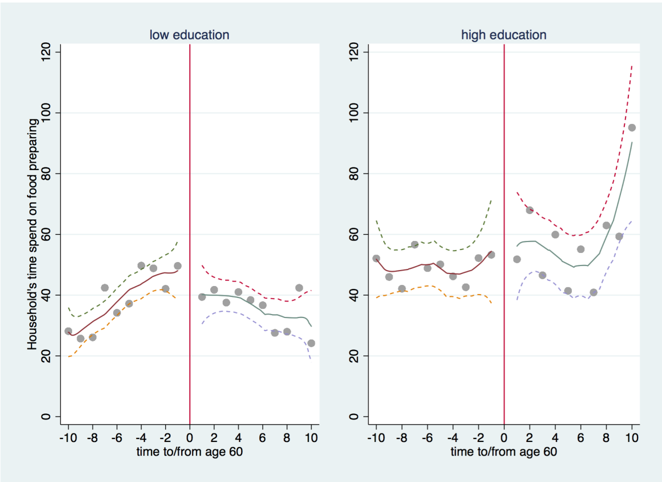

(b) Food shopping time of the household

Figure 3.10: Mechanism: Food Shopping Time in the CHNS Sample

Note: This graph is analysed based on the CHNS dataset. Cells are the household head in each age from age 50 to 70. Panel (a) shows the food shopping time pattern of the household head only for low educated group (left panel) and high educated group (right panel). Panel (b) shows the household food shopping time pattern for low educated group (left panel) and high educated group (right panel).




(a) Food preparing time of the individual




(b) Food preparing time of the household

Figure 3.11: Mechanism: Food Preparing Time in the CHNS Sample

Note: This graph is analysed based on the CHNS dataset. Cells are the household head in each age from age 50 to 70. Panel (a) shows the food preparing time pattern of the household head only for low educated group (left panel) and high educated group (right panel). Panel (b) shows the household food preparing time pattern for low educated group (left panel) and high educated group (right panel).

# References

Adam, S. and M. Brewer (2005). Take-up of family credit and working families' tax credit: final report.

Aguiar, M. and E. Hurst (2005). Consumption versus. expenditure. *Journal of Political Economy 113*(5), 919–948.

Aguiar, M. and E. Hurst (2013). Deconstructing lifecycle expenditure. *Journal of Political Economy 121*(3), 437–492.

Aguila, E., O. Attanasio, and C. Meghir (2011). Changes in consumption at retirement: evidence from panel data. *Review of Economics and Statistics 93*(3), 1094–1099.

Almond, D., H. Li, and S. Zhang (2013). Land reform and sex selection in china. *NBER Working Paper No. 19153*.

Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc (2009). Do value-added estimates add value? accounting for learning dynamics. *Harvard Kennedy School of Government Working Paper* (4435671).

Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc (2011). Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics 3*(3), 29–54.

Angrist, J., V. Lavy, and A. Schlosser (2010). Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics 28*(4), 773–824.

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on student achievement. *Quarterly Journal of Economics 114*(2), 533–75.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Banerjee, A., X. Meng, and N. Qian (2010). The life cycle model and household savings: Micro evidence from urban china. *National Bureau of Economic Research*.

Banks, J., R. Blundell, and S. Tanner (1998). Is there a retirement-savings puzzle? *American Economic Review 88*(4), 769–788.

Barrett, G. F. and M. Brzozowski (2012). Food expenditure and involuntary retirement: Resolving the retirement-consumption puzzle. *American Journal of Agricultural Economics 94*(4), 945–955.

Battistin, Erich, A. B. E. R. and G. Weber (2009). The retirement consumption puzzle: Evidence from a regression discontinuity approach. *American Economic Review*, 2209–2226.

Becker, G. S. (1965). A theory of the allocation of time. *The economic journal 75*(299), 493–517.

Becker, G. S. . and N. Tomes (1976). Child endowments and the quantity and quality of children. *Journal of Political Economy 84*(4), S143–62.

Becker, G. S. and T. Nigel (1973). On the interaction between quantity and quality of children. *Journal of Political Economy 81*(2), 279–288.

Becker, G. S. and T. Nigel (1976). Child endowments and the quantity and quality of children. *Journal of Political Economy 84*(4), 143–162.

Bernal, R. (2008). The effect of maternal employment and child care on children's cognitive development. *International Economic Review 49*(4), 1173–1209.

Bernal, R. and M. P. Keane (2010). Quasi-structural estimation of a model of childcare choices and child cognitive ability production. *Journal of Econometrics 156*(1), 164–189.

Bernal, R. and M. P. Keane (2011). Child care choices and children's cognitive achievement: The case of single mothers. *Journal of Labor Economics 29*(3), 459–512.

Bernheim, B. D., J. Skinner, and S. Weinberg (2001). What accounts for the variation in retirement wealth among us households? *American Economic Review*, 832–857.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2005). The more the merrier? the effect of family size and birth order on children's education. *The Quarterly Journal of Economics*, 669–700.

Blau, D. M. (1999). The effect of income on child development. *Review of Economics and Statistics 81*(2), 261–276.

Blundell, R., M. Brewer, and M. Francesconi (2008). Job changes and hours changes: understanding the path of labor supply adjustment. *Journal of Labor Economics 26*(3), 421–453.

Blundell, R., M. Brewer, and A. Shephard (2005). Evaluating the labour market impact of working families' tax credit using difference-in-differences.

Brewer, M., M. Francesconi, P. Gregg, and J. Grogger (2009). Feature: In-work benefit reform in a cross-national perspective-introduction. *The Economic Journal 119*(535), F1–F14.

Browning, M., T. F. Crossley, and M. Lührmann (2014). Durable purchases over the later life cycle. *Oxford Bulletin of Economics and Statistics 78*(2), 145–169.

Chamon, M. D. and E. S. Prasad (2010). Why are saving rates of urban households in china rising? *American Economic Journal: Macroeconomics*, 93–130.

Chiappori, P.-A. (1992). Collective labor supply and welfare. *Journal of political Economy*, 437–467.

Clark-Kauffman, E., G. J. Duncan, and P. Morris (2003). How welfare policies affect child and adolescent achievement. *The American Economic Review 93*(2), 299–303.

Dahl, G. B. and L. Lochner (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *The American Economic Review 102*(5), 1927–1956.

Das Gupta, M., J. Zhenghua, L. Bohua, X. Zhenming, W. Chung, and B. Hwa-Ok (2003). Why is son preference so persistent in east and south asia? a cross-country study of china, india and the republic of korea. *The Journal of Development Studies 40*(2), 153–187.

Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications.

Deaton, A. and J. Muellbauer (1980). An almost ideal demand system. *The American economic review*, 312–326.

Del Bono, E., M. Francesconi, Y. Kelly, and A. Sacker (2016). Early maternal time investment and early child outcomes. *The Economic Journal 126*(596).

Dominitz, J. (1998). Earnings expectations, revisions, and realizations. *Review of Economics and statistics 80*(3), 374–388.

Dong, Y. and D. T. Yang (2016). Mandatory retirement and the consumption puzzle: Prices decline or quantities decline? *SSRN 2744760*.

Dooley, M. and J. Stewart (2004). Family income and child outcomes in canada. *Canadian Journal of Economics/Revue canadienne d'économique 37*(4), 898–917.

Duncan, G. J., P. L. Chase-Lansdale, et al. (2000). *Welfare reform and child well-being.* Citeseer.

Duncan, G. J., W. J. Yeung, J. Brooks-Gunn, and J. R. Smith (1998). How much does childhood poverty affect the life chances of children? *American sociological review*, 406–423.

Eissa, N. and J. B. Liebman (1996). Labor supply response to the earned income tax credit. *The Quarterly Journal of Economics 111*(2), 605–637.

Fisher, L., D. Dixon, J. Herson, R. Frankowski, M. Hearron, and K. E. Peace (1989). Intention to treat in clinical trials.

Fisher, P. (2016). British tax credit simplification, the intra-household distribution of income and family consumption. *Oxford Economic Papers 68*(2), 444–464.

Francesconi, M., H. Rainer, and W. Van der Klaauw (2008). Unintended consequences of welfare reform: the case of divorced parents.

Francesconi, M. and W. Van der Klaauw (2007). The socioeconomic consequences of "in-work" benefit reform for british lone mothers. *Journal of Human Resources 42*(1), 1–31.

Gelman, A. and G. Imbens (2014). Why high-order polynomials should not be used in regression discontinuity designs. Technical report, National Bureau of Economic Research.

Giles, J., D. Wang, and C. Zhao (2010). Can chinaâĂŹs rural elderly count on support from adult children? implications of rural-to-urban migration. *Journal of Population Ageing 3*(3-4), 183–204.

Gregg, P., S. Harkness, and S. Smith (2009). Welfare reform and lone parents in the uk. *The economic journal 119*(535), F38–F65.

Grogger, J. and L. A. Karoly (2009). The effects of work-conditioned transfers on marriage and child well-being: A review. *The Economic Journal 119*(535), F15–F37.

Grogger, J., L. A. Karoly, and J. Grogger (2009). *Welfare reform: Effects of a decade of change.* Harvard University Press.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Haider, S. J. and M. Stephens (2007). Is there a retirement-consumption puzzle? evidence using subjective retirement expectations. *The review of economics and statistics 89*(2), 247–264.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science 312*(5782), 1900–1902.

Huang, W., X. Lei, and Y. Zhao (2016). One-child policy and the rise of man-made twins. *Review of Economics and Statistics 98*(3), 467âĂŞ476.

Hurst, E. (2008). The retirement of a consumption puzzle. *National Bureau of Economic Research No. w13789*.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics 142*(2), 615–635.

Lee, D. S. and D. Card (2008). Regression discontinuity inference with specification error. *Journal of Econometrics 142*(2), 655–674.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Lee, J. (2008). Sibling size and investment in children's education: An asian instrument. *Journal of Population Economics 21*(4), 855–875.

Lei, X., J. Strauss, M. Tian, and Y. Zhao (2015). Living arrangements of the elderly in china: evidence from the charls national baseline. *China economic journal 8*(3), 191–214.

Li, H., X. Shi, B. Wu, et al. (2015). The retirement consumption puzzle in china. *American Economic Review papers and proceedings 105*(5), 437–41.

Li, H., J. Zhang, and Y. Zhu (2008). The quantity-quality trade-off of children in a developing country: Identification using chinese twins. *Demography 45*(1), 223–243.

Li, Z. (2012). The basic old-age insurance of china: Challenges and countermeasures. *World Pension Summit*.

Lin, M.-J., N. Qian, and J.-T. Liu (2008). More women missing, fewer girls dying: The impact of abortion on sex ratios at birth and excess female mortality in taiwan. *National Bureau of Economic Research No. 14541*.

Liu, H. (2014). The quality–quantity trade-off: evidence from the relaxation of china's one-child policy. *Journal of Population Economics 27*(2), 565–602.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

McElroy, M. and D. T. Yang (2000). Carrots and sticks: fertility effects of china's population policies. *The American economic review 90*(2), 389–392.

Meng, X. and R. Gregory (2007). Exploring the impact of interrupted education on earnings: the educational cost of the chinese cultural revolution.

Miller, A. R. and L. Zhang (2009). The effects of welfare reform on the academic performance of children in low-income households. *Journal of policy Analysis and Management 28*(4), 577–599.

Milligan, K. and M. Stabile (2011). Do child tax benefits affect the well-being of children? evidence from canadian child benefit expansions. *American Economic Journal: Economic Policy 3*(3), 175–205.

Modigliani, F. and R. Brumberg (1954). Utility analysis and the consumption function: An interpretation of cross-section data. *Franco Modigliani 1*.

Modigliani, F. and S. L. Cao (2004). The chinese saving puzzle and the life-cycle hypothesis. *Journal of economic literature 42*(1), 145–170.

Mogstad, M. and M. Wiswall (2016). Testing the quantity–quality model of fertility: Estimation using unrestricted family size models. *Quantitative Economics 7*(1), 157–192.

Qian, N. (2009). Quantity-quality and the one child policy: The only-child disadvantage in school enrollment in rural china. *National Bureau of Economic Research No. 14973*.

Rodriguez, G. (2007). Lecture notes on generalized linear models. *http://data.princeton.edu/wws509/notes/ 4*.

Rosenzweig, M. R. and K. I. Wolpin (1980). Testing the quantity-quality fertility model: The use of twins as a natural experiment. *Econometrica*, 227–240.

Rosenzweig, M. R. and J. Zhang (2009). Do population control policies induce more human capital investment? twins, birth weight and china's "one-child" policy. *The Review of Economic Studies 76*(3), 1149–1174.

Schultz, T. P. (2007). Population policies, fertility, women's human capital, and child quality. *Handbook of development economics 4*, 3249–3303.

Shonkoff, J. P., D. A. Phillips, et al. (2000). *From neurons to neighborhoods: The science of early childhood development.* National Academies Press.

Silva, J. S., S. Tenreyro, and K. Wei (2014). Estimating the extensive margin of trade. *Journal of International Economics 93*(1), 67–75.

Smith, S. (2006). The retirement-consumption puzzle and involuntary early retirement: Evidence from the british household panel survey. *The Economic Journal 116*(510), C130–C148.

Stancanelli, E. G. and A. Van Soest (2012). Retirement and home production: A regression discontinuity approach. *American Economic Review papers and proceedings 102*(3), 600–605.

Stephens Jr, M. (2004). Job loss expectations, realizations, and household consumption behavior. *Review of Economics and statistics 86*(1), 253–269.

Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology 51*(6), 309.

Todd, P. E. and K. I. Wolpin (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal 113*(485), F3–F33.

Todd, P. E. and K. I. Wolpin (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human capital 1*(1), 91–136.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *International Economic Review 43*(4), 1249–1287.

Wang, W. and F. Famoye (1997). Modeling household fertility decisions with generalized poisson regression. *Journal of Population Economics 10*(3), 273–283.

Wei, S.-J. and X. Zhang (2011). The competitive saving motive: Evidence from rising sex ratios and savings rates in china. *Journal of Political Economy 119*(3), 511–564.

Willis, R. J. (1973). A new approach to the economic theory of fertility behavior. *Journal of political Economy 81*(2), S14–S64.

Windmeijer, F. A. and J. M. Santos Silva (1997). Endogeneity in count data models: an application to demand for health care. *Journal of applied econometrics 12*(3), 281–294.

Winkelmann, R. and K. F. Zimmermann (1994). Count data models for demographic data. *Mathematical Population Studies 4*(3), 205–221.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.

Yang, D. T. (2012). Aggregate savings and external imbalances in china. *The Journal of Economic Perspectives 26*(4), 125–146.

Yi, Z., T. Ping, G. Baochang, X. Yi, L. Bohua, and L. Yongpiing (1993). Causes and implications of the recent increase in the reported sex ratio at birth in china. *Population and Development Review*, 283–302.

Zaslow, M. J., K. A. Moore, J. L. Brooks, P. A. Morris, K. Tout, Z. A. Redd, and C. A. Emig (2002). Experimental studies of welfare reform and children. *The future of children*, 79–95.