NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS
International College of Economics and Finance

Danila Karapsin

The effect of family size on educational attainment: case of Russia

Term paper

38.04.01 ECONOMICS

Master's Programme **'Financial Economics'**

Contents

1   Introduction

World total fertility rates have been declining monotonically for many decades and are now below the replacement level for many countries. That raises questions about the mechanisms of that dynamic and the factors that affect it. However, modern economics has answers for it. The economic theory of fertility, which emerged in Becker (1960) and developed in Becker and Lewis (1973) and Becker and Tomes (1976), provides a foundation for macroeconomic models of general equilibrium with endogenous fertility; see, for example, Becker, Murphy and Tamura (1990). It also explains, from an economic point of view, the dynamic of the demographic transition and predicts the U-shape relationship between income and fertility. Moreover, it claims that at the optimum, there is a tradeoff between the quantity of children and their quality.

Like any theory, that one needs to be checked empirically. One popular way to do it is to check if a quantity-quality tradeoff really exists. Dozens of papers attempted to do that, for example, by trying to apply a quasi-experimental approach to that problem. Rosenzweig and Wolpin (1980) pioneered the use of twins for instrumental variable construction. Later, Rosenzweig and Zhang (2009) showed that such an instrument might be internally invalid due to resource reallocation within the family, as predicted by Becker and Tomes (1976). It has been shown that in China, parents prefer to reinforce differences in quality endowments, which creates a positive bias for IV-based estimates. In contrast, Black et al. (2010) provided evidence for compensation for endowment differences in Norway. Using the same data, Mogstad and Wiswall (2016) concluded that fertility policy in Norway reduces the costs of having one child more than the costs of having two or more children.

As can be seen, on the way to evidence for the quantity-quality tradeoff, many positive externalities can be created. However, there are no papers that are trying to look for it in Russia. This study aims to fill this gap.

2 Literature review

2.1 Theoretical models

Family size

Although the interest of economists in human fertility can be traced back to Malthusianism, the modern approach to that topic started in the postwar years of the 20th century, when the extrapolations based on the macro data suddenly stopped to deliver accurate demographic predictions. One of the most well-known frameworks originates from Becker (1960). In the proposed model (which is now known as the Quantity-Quality (QQ) model), kids are viewed as a consumer good; however, not only the number of children (n) brings utility to the family, but also their quality (q, which is assumed to be equal across all kids within the family) and consumption of other goods (c). That, together with the uncertainty in the children's production, budget constraint and preferences of the family, gives rise to the problem of expected utility maximization. The most relevant point from that model for our study is that the assumption about substitutability between quality and quantity of kids leads to the substitution of quantity with quality if income elasticity for the quality is higher than for the quantity.

Later in Becker and Lewis (1973), the following formalization of that problem was analyzed:

$$\begin{cases} U = f(n, q, c) \to \max\limits_{n \geqslant 0, q \geqslant 0, c \geqslant 0} \\ I = nqp_q + cp_c \end{cases}$$

It turned out that in the proposed framework, quantity and quality of children are closely related in the sense that an increase in quality also increases the marginal costs of the quantity; the converse is also true. That conclusion holds without any additional assumptions about elasticities or substitutability because of the budget constraint's form (particularly because of the interaction between n and q). The proposed model not only explains the demographic transition dynamic but also implies that a higher quantity of kids leads to lower quality.

The following paper on this topic by Becker and Tomes (1976) generalized these findings by introducing endowments of quality for the kids (which depend on, e.g., inherited ability, government support, etc.), which may differ across them. Such a setup allows q to vary within one family, but the quantity-quality tradeoff still persists. One important prediction is that, in response to the endowment differences across children, parents may reallocate the resources of the family in order to achieve the desired optimum. Depending on the preferences of the family, it might be that

more resources will be allocated to the more gifted kids (reinforcement), or vice versa, kids with lower endowment will receive additional support (compensation).

The further development of the fertility theory provided by Becker and Barro (1988) establishes a link between fertility and economic growth. In the suggested model, there is no kid's 'quality'; however, parents care about the utilities of their children, i.e., they are maximizing not only their own utility but also the utility of their offspring; they, in turn, care about their descendants, and so on, which results in dynastic utility. The QQ tradeoff here manifests in the form of a negative association between the costs of having a child (i.e., how much the utility of that child costs) and the quantity of kids in the optimum. This paper started a whole stream of literature devoted to the issues related to endogenous fertility and human capital accumulation in general equilibrium models. A comprehensive review is provided in Guo et al. (2022).

Birth order

Many empirical papers related to the issues of interactions between fertility and the quality of children also use a kid's birth order as the control variable or even the variable of interest. As with family size, the birth order and its effect on a variety of outcomes (not only on the child's quality) have been a subject of interest to scientists from various fields (not only economics but also psychology, medicine, etc.) for centuries. However, in this case, there is no such long history of development of the theoretical models related to the birth order; most papers are empirical, partially because their main contributors are from those sciences whose theoretical side is more closely tied to their empirics (e.g., psychology) than in economics.

One of the most influential frameworks relevant for this study was proposed by Zajonc and Markus (1975) and is now known as the 'Confluence Model.' The idea is that in every family, every individual has some degree of intellectual ability, which is higher for the adults and lower for the children. Then, the firstborns have an advantage over their younger siblings since, for some time, they are growing in an environment where the average ability is higher. And that principle continues: secondborns have an advantage over their younger siblings, and so on. Another explanation known as the 'Resource Dilution' is provided in Blake (1981). The idea is very similar, though: every family has a finite number of resources (money, parental attention, etc.) from which kids may benefit. Then, very similarly to the 'Confluence model,' firstborns extract more advantages from that situation simply because, for some time, the resource flow is not diluted; laterborns, however, start to receive an already diluted fraction of the available resources. It is worth noticing that both approaches do not refute QQ theory or contradict its conclusions, but rather present a different way to derive the same conclusions about the quality-quantity relation-

ship. Indeed, if we accept that a higher amount of benefits provides a higher quality of the child (one of the assumptions in the QQ theory), then, ceteris paribus, the firstborn in the family with one child would have a higher quality than the firstborn from the family with two children, since the dilution of the resource flow decreases the quality of the firstborn (because lower input implies lower quality). Furthermore, every firstborn has a positive probability of being the only kid in the family their whole life; secondborns, however, always end up in families with at least two kids, and so on.

2.2   Empirical findings

Since one of the main pillars of the QQ theory is the QQ tradeoff, the most popular way to test it nowadays is to check if there is any significant negative relationship between a kid's quality and the quantity of children in the related family. Then, however, a question about measuring the child's quality and endogen eity removal arises. For the proxy of a kid's quality, some form of educational outcome is often chosen, which is quite reasonable: the more education an individual has, the more value he may bring to the production process, which corresponds to the definition of human capital. There is another group of studies that use earnings as a measure of quality. However, in this paper, they are mostly ignored; the main reason is the limited data availability. One would need data on individuals' families and earnings after at least 10 years after they had finished their education to estimate that properly. For now, there is no such data for Russia.

The challenges that accompany determining the birth order causal inference were described in Blake (1989). Firstly, family size has to be used as the control variable; otherwise, birth order and family size effects will be confounded. Secondly, control for the birth years of kids and parents is needed because laterborns may enjoy the benefits of technological progress that were not developed while the firstborn, for example, was growing up. Controlling for birth spacing is also proposed; however, Black et al. (2005) argue that, claiming the endogenous nature of this variable, it simply might be optimally chosen by families. Furthermore, all papers that will be described further found little or no effect of birth spacing on quality.

One of the pioneering studies was done by Rosenzweig and Wolpin (1980) on household data from India. In this paper, the impact of family size on the schooling index (developed by the authors) had been analyzed. It is the first paper in which researchers used a quasi-experimental approach in order to avoid the endogeneity of

family size; the twin births were used as the instrument for the quantity of kids. The idea is that while family size is endogenously chosen by parents and that choice is defined by variables that also may affect quality outcomes, twin occurrence can be considered a random event that is uncorrelated with such variables. The negative relationship between schooling attainment and family size had been found. However, in the following decades, almost no studies used such an instrument, probably because of the data shortage.

One of the first papers to investigate the QQ tradeoff was Butcher and Case (1994). Here, US data from the famous PSID, NLSW, and CPS surveys had been utilized in order to estimate the effect sibship's gender composition on the women's educational attainment (measured as the years of schooling) and earnings. Among other results, a robust and significant negative impact of the sibship's size on educational attainment for women (but not for men) had been found. In contrast, a Kaestner (1994) study based on US data from the NLSY found little or no effect of sibship's gender composition on educational attainment; the similar conclusion had been derived by Hauser and Kuo (1998), based on US data from OCG, SIPP, and NSFH surveys. Morduch (2000) found a significant impact of gender composition on educational attainment in Tanzania but not in South Africa. Such controversial results are often used as an argument against the usage of the gender composition variables as an instrument (like it is used in, for example, Conley (2004)) for family size (Black et al. (2005), Guo et al. (2022)). Moreover, Black et al. (2010) claim that gender composition instrument estimates effects of the expected fertility increase on child's quality, which is actually irrelevant (detailed explanation in provided further). Because of that and lack of evidence for specific kids' gender preferences in Russia, papers that use such instrument are mostly ignored in the following review.

In Hanushek (1992), data from the Gary Income Maintenance Experiment was used to test the presence of the QQ tradeoff. For the quality measure, test scores were chosen. A significant negative impact of family size on the test scores had been found, while the birth order had no significant effect if family size was controlled. One relevant innovation in that study is the introduction of the education production function:

$$A_{it} = \Phi(F_{it}, S_{it}, X_{it})$$

Where A_{it} is an educational attainment at time t for individual i, F_{it} -family inputs; S_{it} - school inputs; X_{it} - other exogenous inputs. Such an approach highlights the need for controlling the educational background and other inputs that may affect educational attainment; otherwise, estimates for family inputs might be biased. This study, however, is often criticized for the smallness and specificity (only low-income

black families are presented) of the used sample. Another piece of evidence for the QQ theory comes from Iacovou (2001), based on the data from the British National Child Development (BNDC) study. Hanushek's production function had been used. To establish evidence for the QQ tradeoff test scores had been used as the quality measure, and variables related to the birth order had been introduced as the control (e.g., 'youngest of 2 children,' 'youngest of 3 children,' 'middle of 3 children,' etc.), along with many other variables related to the family background. In the end, a significant negative impact of both family size and birth order was found. Unlike Hanushek's paper, the sample used by Iacovou had a much greater size and represented various population groups. However, the high attrition rate (~50%) caused criticism in the following papers.

The next milestone in the empirical studies of the QQ tradeoff was achieved in the paper by Black et al. (2005). Once again, among other approaches, the twin indicator as an instrument for family size was used. That paper pioneered the usage of administrative data in research related to QQ tradeoff identification; their data consists of records that include the entire population of Norway who were aged 16–74 at some point during the 1986–2000 interval. Such luxury not only makes the used sample much greater than any other one that had been used before, but also allows authors to use additional restrictions on the sample data to derive more reliable results. For example, when applying the twin-based instrument for family size, authors exclude twins from the sample since they might differ from the singletone kids in unpredictable ways. Furthermore, the effects are studied only on kids who were born before their twin siblings. In the end, a negative and significant impact of family size on the years of schooling is present only until the birth order is controlled or 2SLS with twins is used.

In the following years, numerous studies utilized a similar approach. In a paper by Cáceres-Delpiano (2006), 5% of the 1980 US Census was used. One of the key assumptions is that any multiple births on the second or higher birth is considered as the exogenous variation in the family size. The sample is restricted to the oldest children, who are not multiple kids. One of the results is that while the negative effect of family size on the probability of private school attendance is not so robust, the negative impact on a mother's labour force participation is consistently negative and significant. It might be evidence of resource reallocation within household which aims to minimize the impact of an unexpected family size increase on the child's quality. In study by Li et al. (2008), the importance of the institutional environment is highlighted: for such countries as Norway, where public education is functioning well and there is significant state support for childbearing, the present family size effects indeed might be small and statistically indistinguishable from 0. However,

in developing countries, families often do not have such benefits; therefore, family size effects might be greater, and it might be easier to detect evidence for the QQ tradeoff. In this study, 1% of the 1990 Chinese Population Census is used. School attendance and education level had been used as the outcome variables. A significant negative impact of family size on the child's quality had been obtained; such effects were quite robust, even after controlling for the birth order and various family characteristics. However, such effects are stronger in rural areas, but in urban China, they decrease greatly or even disappear. In contrast, Qian (2009) found a positive effect of family size on school enrollment using data from the 1% sample of the 1990 China population census matched with the CHNS survey. The relaxation of the One Child Policy had been used as the source of the exogenous variations in family size. Various explanations for that phenomenon were suggested, e.g., economies of scale for schooling might exist. Angrist et al. (2010) used Israeli administrative data and various variables such as twin occurrence, gender composition, and racial difference to construct an instrumental variable. In the end, no significant and robust effects of family size or birth order on the different educational attainment characteristics were found. The study conducted by Åslund and Grönqvist (2010) on the administrative data from Sweden also found little or no effects of family size on different education outcomes.

The study conducted by Rosenzweig and Zhang (2009) is relevant for almost all papers that use the twin instrument. It highlights some key flaws in such a strategy. For example, maternal age needs to be controlled since it is correlated with twinning probability and investments in children. That, however, had been taken into account in most of the of the papers described above. What is usually missed is that twins usually have a lower weight at birth. Numerous studies have shown that birthweight is a significant predictor of earnings, schooling, and other adult outcomes. That might be evidence of the lower quality endowment for the twins; then, as Becker and Tomes (1976) predict, parents may respond with reinforcement or compensation, which will affect estimates of the family size effects for kids who were born earlier. The data from the Chinese Child Twin Survey allowed to estimate how the birthweights of sibling i and sibling j affect the educational input of sibling i, which had been measured as the sum of expenses on tuition, books, educational fees, etc. To take care of the family's fixed effects, cross-sibling difference had been applied. The resulting estimates indicate reinforcement, which implies a positive bias for estimates of family size's effect on the quality of kids who were born before twins. Furthermore, evidence of scale economies had been found for households with twins and households with siblings of the same gender. Such externalities may potentially dilute QQ effects, providing a positive bias for its estimates. In the end, negative

QQ effects were found.

The paper by Black et al. (2010) once again studied the effect of family size on quality using Norway's administrative data and twinning. That time IQ had been used as a measure of the kid's quality, and birthweights had been taken into account. It is worth noting that evidence of compensation rather than reinforcement was found; however, in the end, a significant negative impact of family size on the IQ was obtained. Another important result is that only unexpected fertility increases (e.g., via twins occurrence) affect kids' quality, which explains why estimates that use gender composition as an instrument may provide non-negative estimates for the effects of family size. The explanation is quite intuitive, though. It is reasonable to conclude that if parents wanted to have n kids, the occurrence of singleton birth at any point can hardly be considered a natural experiment because the difference between the desired and obtained quantity or quality does not occur. The family still may achieve an optimum in terms of the quantity and quality of its kids, which was desired before that birth (budget constraint does not bind), and gender's randomness does not play any role here since all uncertainty of the kid's production has already been taken into account by rational agents.

Another important point comes from Mogstad and Wiswall (2016), who used the same Norwegian data, however allowing for non-linearity of the family size effects. The resulting estimates provide positive QQ effects for small families and negative QQ effects for large families. One possible explanation is that fertility-related policies reduce the costs of having one child more than the costs of having two or more children.

3   Data and methodology

3.1   Data

This study is based on the data from the Russian Longitudinal Monitoring Survey - Higher School of Economics (RLMS-HSE), which is a series of yearly nationally representative surveys of Russian households and the individuals from which they consist. Currently, this survey consists of 27 rounds from 1994 to 2022 (there were no surveys in 1997 and 1999). These surveys contain a variety of questions about education and family ties within the household. Unfortunately, sometimes participants of the survey may provide contradictory answers, e. g. participant can report different own birth years in different rounds. Furthermore, not all variables can be found in the survey, most of them have to be constructed. In order to overcome the issues described, the approach that will be described in this section has been developed.

For the birth year of an individual, the minimum of all reported years is used (variables *j69.9c* and *h6*). If in some rounds an individual did not report birth year, it is estimated as the difference between the individual's age and year in which the survey has been conducted (variables age and year). Because it is possible to have non-linear effects, it is reasonable to use that variable as a categorical. However, because of the relatively small sample size, we also want to group birth years into 5-year periods from 1970 to 2005.

An individual is considered a graduate if at least one variable related to his education implies that (variables *educ*, *diplom* and *diplom_1*) or if an individual is a student in a university (variable *j90* = 16) and at least 22 years old (*age*). If an individual had been considered a graduate in one year according to the described rule, he is considered a graduate in all subsequent years. Otherwise, we assume that this individual has no higher education.

For some descriptive statistics, which will be provided later, classification of individuals' jobs is needed. If variable *occup08* takes values from 1 to 5, it is assumed that an individual is working on a job that generally requires higher education. If an individual reports that he is unemployed or there is no information about his employment, he is considered an unemployed person. Otherwise, it is assumed that individuals' jobs do not require higher education.

If the variable *status* is equal to 3 or 4, for some household it is assumed that the family lives in a village. If the variable *popul* takes values between 3 and 7 millions, it can be concluded that the family lives in Saint-Petersburg. If *popul* is higher than 8 million, we may conclude that the family lives in Moscow. If *popul* is higher than

1 million, it is assumed that a family lives in a city with over a million people. In all other cases, a place of residence is defined as a 'city.'

Only those participants for whom we can find both of their parents among other participants are needed for this paper. This restriction limits the size of the sample greatly. Furthermore, note that despite the intention declared by the RLMS-HSE team to keep track of individuals who moved from their original household, it is, obviously, not always possible, so we have some form of the artrition bias here, which leads to distortions, which will be described later. Family ties had been defined from the household-related data. Strictly speaking, that is not always 'family ties,' since in some households individuals are living with 'non-relatives' (value 14 from the codebook for family ties). Family ties are defined by the great variety of variables in the file for households. If a household consists only of kids and parents, it is classified as a nuclear family. For some reasons (probably due to filling mistakes), three individuals had the same gender parents; those individuals were excluded from the sample. If for some individual it is not possible to define family ties within some year, family ties from the closest preceding year are used. For some households, family ties are not completely described, e.g., there is a family with 2 kids and parents; however, according to the derived family ties data, one kid has one sibling and 2 parents, and the other one has only 2 parents. Because of that, the number of kids in each family is defined as the maximum of three values: the number of siblings in the longest family ties array plus 1, the number of unique IDs related to the same parents, and the and the maximal birth order in the family. For one family, it was impossible to define whether grandmother lives within it; this family was excluded from the sample.

Birth order is defined by the ranking of birth years for individuals with the same father and mother IDs. If the birth years of two individuals within one family are the same, they will have the same birth order. If the variable *id_h* is different for a kid and both of the kids' parents in a year X, which is preceding the year Y, such a kid is considered to be separated in the year Y. Finally, from the data about family ties, we can define gender for all kids in the family and check if there are any siblings with different genders for any chosen individual.

In the end, for each family, we truncate the number of kids and birth orders. If a family has 4 or more kids, it is labeled as '>=4'; if the birth order is 3 or higher, it is labeled as '>=3.' Such truncation is needed because of the current demographic situation; the number of observations decreases greatly with an increase in the number of kids. Because the number of individuals from families with 3 kids is greater than the number of individuals who were born third, we can allow later truncation at 4 for the number of kids. We also need only those individuals who are

older than 22; it is assumed that at this point in life, educational attainment is more or less stable, at least if we are talking about higher education. It is also reasonable to limit our sample only to those who were born in 1970 or later, since we do not want to study people whose decisions were affected by events of the very distant past (without that condition, there are people who were born in the 1940s). Finally, it is natural to consider only the earliest observations of individuals that satisfy all the described criteria.

From the file, which contains variables constructed by the RLMS-HSE team, we can extract the expenses of the household (variable *totexpr*). However, for the goals of this study, we need expenses that some household had at the moment before an individual made a decision about higher education, i.e. when an individual was 18 or younger years old. But together with the described age restriction that implies that an individual participated in at least five rounds of the survey or even more, which is not always the case. That limits the number of individuals for whom we can define the lagged household expenses. When we can define lagged expenses for the household, we are estimating to which quartile of the income distribution it belongs (within the year from which we derived these expenses, the distribution is evaluated via a full RLMS sample).

3.2 Variables and descriptive statistics

univ_diploma

This is the dependent variable of that study used to assess an individual's educational attainment; it takes values 0 and 1 (binomial). The alternatives to it are years of schooling and multinomial variable for educational attainment (e.g., with levels '9 classes,' '11 classes'…), but for a number of reasons, the chosen variable is superior to the described alternatives in our case. In the current Russian institutional environment, an individual has many options to construct an educational track if, for example, a doctoral degree is desired, from the longest one, which is: 11 classes, professional education (4 years usually), bachelor and master degrees (+6 years), and only after this 21 years start pursuing the desired degree. The other option is to finish 11 classes, then study 5 years for the higher education degree, and only after this 16 years start pursuing the doctoral degree. As can be seen, in such a situation, years of schooling do not make any sense. The objection against multinomial variable is mostly situational: we have a very limited number of observations, and such fractioning would reduce the possibility of obtaining statistically significant results.

kids_num

That is the variable of interest, so we are especially interested in how it affects the probability of obtaining higher education. This variable is categorical in order to capture possible non-linearity. Note that since family size is endogenously chosen by parents, there might be factors that affect both family size and educational attainment, i.e., there is a possibility of omitted variables bias. Following Hanushek (1992) and Iacovou (2001) we are trying to include controls for family background (which will be described further) in order to avoid omitted variables bias. Based on the reviewed literature, we should expect one of the following results:

1. Family size effects are significant only until we include birth order as the control.

2. Negative and significant family size effects.

kid_order

Following Black et al. (2005) we include birth order into the model. However due to the smallness of the sample and relatively low number of kids in Russian family, it is natural to expect little or no effect of birth order on the probability of obtaining higher education.

residence

Living in a densely populated city not only reduces the costs of obtaining a higher education (e.g., by reducing living costs, easier entry exam preparation, etc.), but also increases the potential benefits of it. Usually, in huge cities, there are more job opportunities, and the quality of education provided by universities is higher. That is why we should expect a positive effect of living in a large city on the probability of obtaining a higher education, while living in a village, for example, should have a negative effect. In such situation 'city' level seems like a natural baseline for that variable.

parents education

Parents are usually trying to ensure that their kids will achieve at least the same social status as they have, and typically, for that, the same level of education is needed. However, the contribution to the child's education from the father and mother might differ. That is why we are including two variables in the model: *father_univ_diploma* and *mother_univ_diploma*.

It is natural to expect that both variables will affect the probability of obtaining a higher education positively, but the effect of *mother_univ_diploma* might be higher due to social standards, which put a lot of kids-related activities on women.

nuclear family

On the one hand, the presence of additional household members may dilute available resources and decrease the probability of obtaining a higher education. On the

Table 1: Income distribution by nuclear family status

label	levels	0	1	Total	p
totexpr	I Q	23 (4.3)	225 (6.7)	248 (6.3)	<0.001
	II Q	60 (11.2)	636 (18.9)	696 (17.8)	
	III Q	144 (26.8)	1102 (32.7)	1246 (31.9)	
	IV Q	310 (57.7)	1411 (41.8)	1721 (44.0)	

Note: p-value of test (chi-square test) for differences between groups is reported.

other hand, the presence of additional adults may increase the income of the household. Such ambiguity is not allowing us to make any predictions about the effect of being a nuclear family on the dependent variable. Note that in our sample share of the families within the highest quartile is significantly higher for non-nuclear families (chi-square test); this result is presented in the following table.

is separated

As has been mentioned earlier, sometimes individuals may move from their original household to another, and sometimes RLMS is able to track that. Potentially, kids who are moving away from their parents may have very different characteristics from those who stayed; these characteristics may also affect their probability of obtaining a higher education. That is the rationale for including that variable in the model. However, it is almost impossible to predict how it affects our dependent variable since separated individuals are prone not only to separation but also to staying in touch with the RLMS, and it is quite complicated to predict the effect of such attrition bias.

birth_year_trunc

Earlier, we divided all individuals's birth years into 5-year periods. It is reasonable to assume that technological progress and the emergence of the new means of communication decrease the costs of higher education and increase benefits from it at the same time, so we should expect that individuals who were born later will be more likely to obtain higher education. This point is illustrated in the following graph, as we can see the share of individuals with higher education grows with time.

other gender siblings

Public beliefs about gender roles may affect resource allocation between siblings of different genders in one family. However, such effects should not be great nor significant in Russian society, which has been quite urbanized since the second half of the previous century; furthermore, there is no evidence of specific gender preferences in Russia.
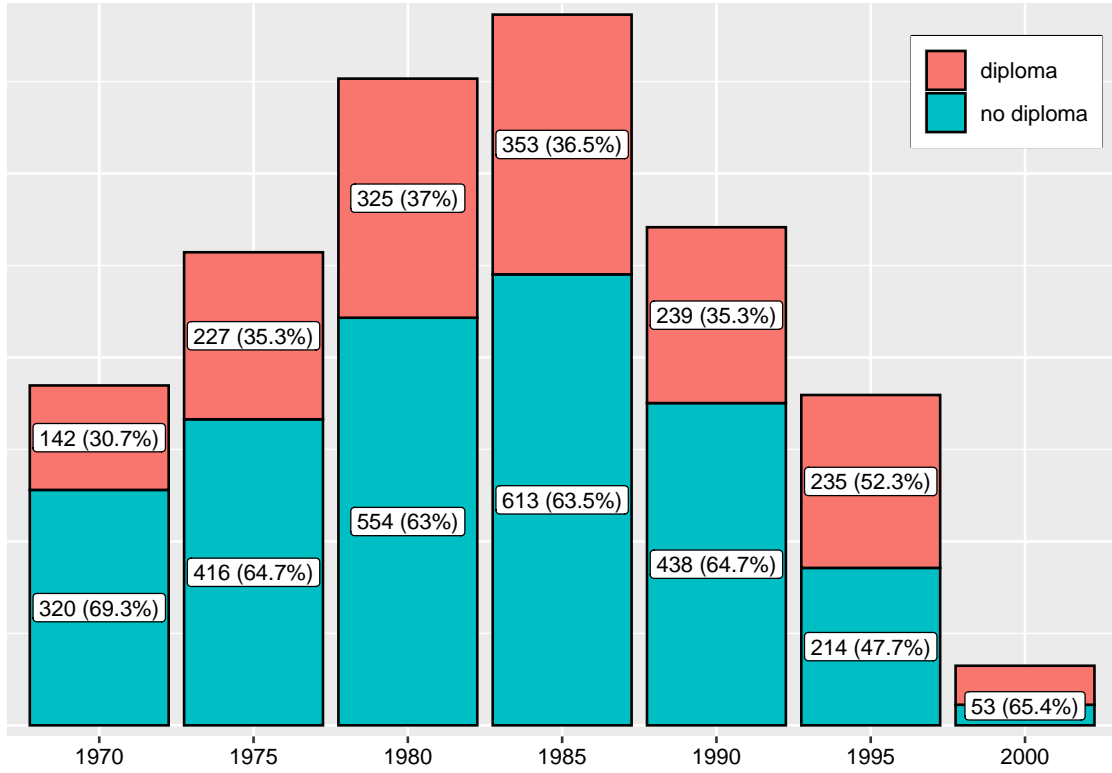
gender

Figure 1: Number of individuals by birth year, gender and diploma status

The relevance of that variable will be shown further.

Descriptive statistics

The resulting sample consists of 4157 individuals. The following table contains some descriptive statistics for the variables that will be used in the model, grouped by gender. For numerical variables, averages and standard errors are reported, while for categorical variables, the number of observations and shares are shown. To assess the significance of differences between males and females, a t-test (for continuous variables) and a chi-square test (for categorical variables) are used. P-values are reported in the 'p' column. Note that the number of individuals who live in Saint-Petersburg is very low; still, we need to consider that city as a separate level for the residence variable since its status is quite different from both Moscow and the rest of Russian cities. As can be seen, for the majority of variables, gender differences are not significant. However, for females, the share of individuals with higher education and separated individuals is significantly greater than for males. In the case of the profession type, we can observe that there is a significant difference, mostly because there are more females who are working on jobs that require higher education.

Table 2: Descriptive statistics by gender

label	levels	female	male	Total	p
age	Mean (SD)	24.0 (3.9)	24.1 (4.0)	24.0 (3.9)	0.464
birth_year	Mean (SD)	1984.9 (7.7)	1984.8 (7.8)	1984.9 (7.7)	0.842
residence	city	911 (46.9)	1028 (46.4)	1939 (46.6)	0.166
	village	617 (31.8)	754 (34.0)	1371 (33.0)	
	1mln	173 (8.9)	155 (7.0)	328 (7.9)	
	MSK	185 (9.5)	214 (9.7)	399 (9.6)	
	SPB	56 (2.9)	64 (2.9)	120 (2.9)	
kids_num	1	763 (39.3)	815 (36.8)	1578 (38.0)	0.186
	2	885 (45.6)	1050 (47.4)	1935 (46.5)	
	3	206 (10.6)	263 (11.9)	469 (11.3)	
	>=4	88 (4.5)	87 (3.9)	175 (4.2)	
kid_order	1	1295 (66.7)	1528 (69.0)	2823 (67.9)	0.231
	2	538 (27.7)	562 (25.4)	1100 (26.5)	
	>=3	109 (5.6)	125 (5.6)	234 (5.6)	
nuclear_family	0	273 (14.1)	307 (13.9)	580 (14.0)	0.890
	1	1669 (85.9)	1908 (86.1)	3577 (86.0)	
parents_education	1 diploma	500 (25.7)	541 (24.4)	1041 (25.0)	0.510
	2 diplomas	319 (16.4)	355 (16.0)	674 (16.2)	
	no diplomas	1123 (57.8)	1319 (59.5)	2442 (58.7)	
lag_expenses	I Q	63 (6.3)	84 (7.5)	147 (6.9)	0.595
	II Q	203 (20.3)	215 (19.2)	418 (19.7)	
	III Q	311 (31.0)	361 (32.2)	672 (31.7)	
	IV Q	425 (42.4)	460 (41.1)	885 (41.7)	
is_separated	0	1731 (89.1)	2098 (94.7)	3829 (92.1)	<0.001
	1	211 (10.9)	117 (5.3)	328 (7.9)	
profession_type	not qualifed	118 (6.1)	827 (37.3)	945 (22.7)	<0.001
	qualified	1081 (55.7)	669 (30.2)	1750 (42.1)	
	unemployed	743 (38.3)	719 (32.5)	1462 (35.2)	
univ_diploma	diploma	871 (44.9)	703 (31.7)	1574 (37.9)	<0.001
	no_diploma	1071 (55.1)	1512 (68.3)	2583 (62.1)	

Note: p-values of tests (t-test and chi-square test) for differences between genders are reported.
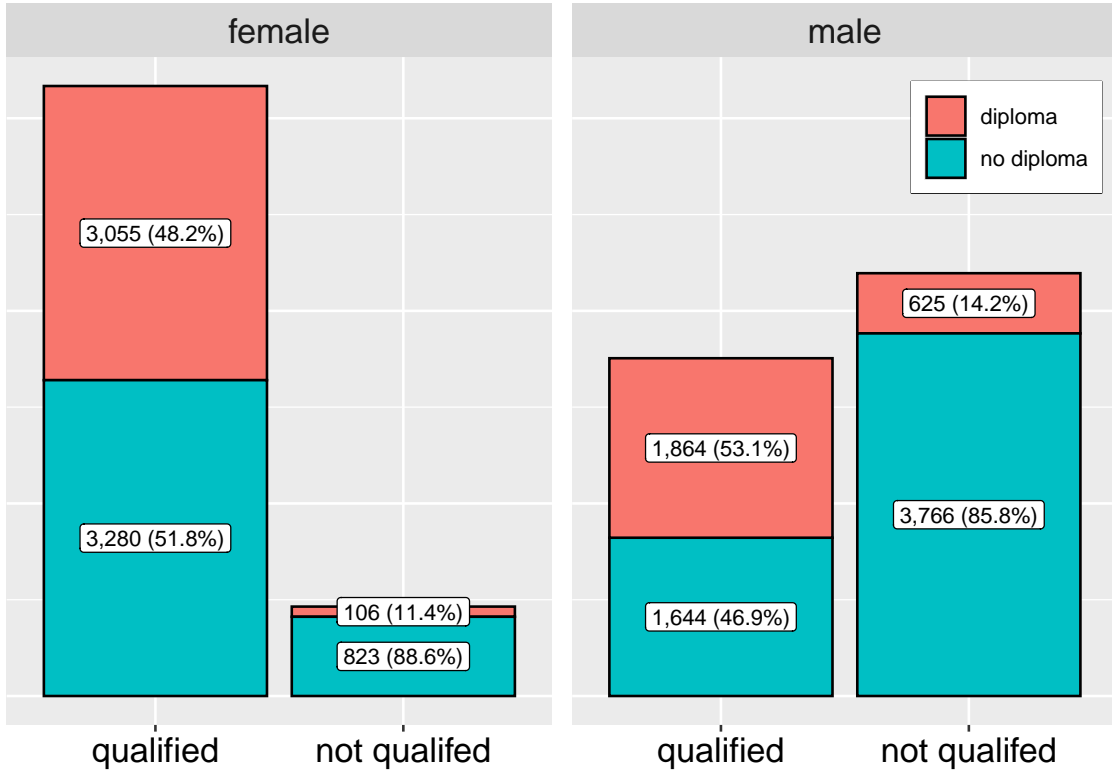
Figure 2: Higher education in professional groups by gender

The latter point may explain the observed difference between gender shares among graduates. Public beliefs about gender roles, together with biological differences, limit the set of jobs from which females can choose. As a result, the value of higher education for women is much higher than for men, since without it, there are fewer job opportunities than for mens available. Figure 1 illustrates that idea using the data from the unfiltered RLMS sample about the number of employed individuals by gender and professional group. There is no straightforward explanation of the female prevalence among the separated individuals; one can think of that as a characterization of the attrition bias that we have (it is easier to keep tracking females than males for some reason).

Another helpful exercise is to assess possible distortions in the sample that we have due to the attrition bias that has been described earlier. That is not generally possible; however, in our specific case for each year, we have a representative sample from RLMS. The idea is the following: we are going to make the same transformation with the full RLMS sample that we have performed for the fraction of individuals for whom we were able to define parents (filtering by age, then selecting the earliest observations for each individual); in other words, we are constructing the new sample as if we are able to define parents for all individuals. That new sample should be, in
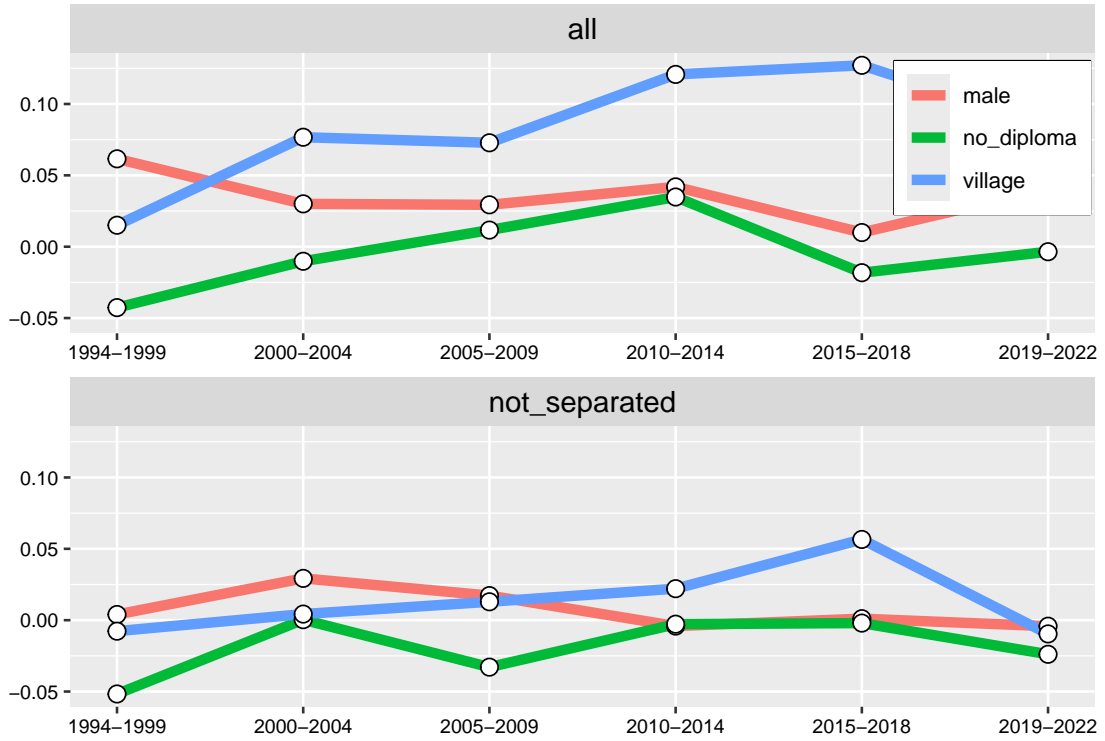
Figure 3: Distortions in the selected sample

some sense, representative. Then, using the representative sample, we can measure deviations within the selected groups in our sample from the representative one, year by year. To provide even more comparability, individuals with an age higher than 26 were excluded for both selected and representative samples. With the same goal in mind, we are excluding those individuals who are living with their parents and for these parents higher education status is unknown (with the approach that has been described earlier, that simply means that these parents never participated in the survey). This procedure has been repeated twice: firstly for all individuals, and secondly only for those who are living with their parents (in both samples). Note that the resulting 'representative' sample does not fully coincide with the one that we obtained earlier since it contains families where only one parent is present.

The results for differences in representation of different groups are plotted on the graph above. Note that for illustrative purposes, average differences for selected time periods are shown, e.g., in 1994–1999, the average share of people without higher education in our sample is higher by 15.7 percentage points than in the representative sample (for not-separated individuals). As can be seen, when we are comparing the shares people who live in a village for all individuals, it can be concluded that our sample overrepresents villagers. This overrepresentation, however, decreases when

19

we are considering only non-separated individuals. One way to explain the observed phenomenon is to conclude that individuals who live in villages are generally less likely to separate from their original household, or that it is less likely that RLMS will be able to track that separation. The key point here is that the provided algorithm for sample formation results in an overrepresentation of people who are not separating from their original households; hence, any conclusion with a claim of generality should take that into account.

3.3 Empirical strategy

To estimate the effect of family size on probability of obtaining higher education we are going to use logit model, i.e., the formal goal of this study is to correctly define α from the following equation:

$$P(\text{UnivDiploma} = 1) = f(\alpha * \text{KidsNum} + \sum_{i=1}^{N} \beta_i * x_i)$$

Where $f(x)$ is:

$$f(x) = \frac{1}{1 - e^{-x}}$$

And each of x_i corresponds to one of the control variables that have been described before. We are going to include these controls one by one, checking estimates and their significance at each step. This procedure will be repeated twice: the first time with the full sample and the second time with the filtered sample, which consists only of not-separated individuals. To get an idea of the quality of the final model, we are going to estimate its AUC scores along with accuracy, precision, and recall multiple times using various proportions for train and test samples from 0.25 to 0.75 with a step equal to 0.01. At each step, sample splitting will be performed 300 times, and then the average value for each metric will be calculated.

4   Preliminary results

This section is devoted to a brief description and discussion of the obtained results. Estimates, standard errors, and p-values are presented in the appendix, as well as the values of quality metrics.

As can be seen, a significant negative impact of family size on the probability of obtaining a higher education has been shown. Both estimates and p-values are quite robust to different specifications; for all models, a negative impact of family size has been obtained. Birth order effects, in contrast with the findings of Black et al. (2005), are relatively small and statistically insignificant, which might be due to the institutional or cultural differences between Norway and Russia. However, most estimates are negative and decrease for later births. The model delivers accuracy that is consistently higher than 0.7, i.e., ~30% of guesses based on it will be incorrect.

One concern about the obtained results is the smallness and specificity (bias towards not-separated individuals) of the sample. The other is the possibility of omitted variable bias. Despite the plenty of control variables that have been included in the model, there still might be a missing piece; furthermore, the probability of obtaining a higher education might be affected by such variables, which are also correlated with family size but at the same time are family-specific and cannot be measured explicitly. One way to solve these problems is to introduce a twin-based instrumental variable for family size. But here once again, the smallness of the sample creates limitations. The promising direction of further research is to utilize the publicly available microdata from the Russian Censuses of 2002 and 2010 in order to construct a twin-based intrumental variable for family size. To get an idea about preferences of Russian families in terms of reinforcement/compensation and, hence, sign of the bias in estimates, RLMS data can be used.

5 List of References

ANGRIST, J. D., AND W. N. EVANS. (1998): "Children and their parents' labor supply: Evidence from exogenous variation in family size," *American Economic Review*, 88, 450–77.

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, 773–824.

ÅSLUND, O., AND H. GRÖNQVIST. (2010): "Family size and child outcomes: Is there really no trade-off?" *Labour Economics*, 17, 130–39.

BECKER, G. S. (1960): "An Economic Analysis of Fertility," in *Demographic and economic change in developed countries*,. NBER chaptersNational Bureau of Economic Research, Inc, 209–40.

BECKER, G. S., AND R. J. BARRO. (1988): "A reformulation of the economic theory of fertility," *The Quarterly Journal of Economics*, 103, 1–25.

BECKER, G. S., AND H. G. LEWIS. (1973): "On the interaction between the quantity and quality of children," *Journal of Political Economy*, 81, S279–88.

BECKER, G. S., AND N. TOMES. (1976): "Child endowments and the quantity and quality of children," *Journal of Political Economy*, 84, S143–62.

BLAKE, J. (1981): "Family size and the quality of children," *Demography*, 18, 421–42.

---. (1989): *Family Size and Achievement*, Berkeley: University of California Press.

BUTCHER, K., AND A. CASE. (1994): "The effect of sibling sex composition on women's education and earnings," *The Quarterly Journal of Economics*, 109, 531–63.

CÁCERES-DELPIANO, J. (2006): "The impacts of family size on investment in child quality," *The Journal of Human Resources*, 41, 738–54.

CONLEY, D. (2004): What Is the "True" Effect of Sibship Size and Birth Order on Education? Instrumental Variable Estimates from Exogenous Variation in Fertility.",mimeo.

DOEPKE, M. (2015): "GARY BECKER ON THE QUANTITY AND QUALITY OF CHILDREN," *Journal of Demographic Economics*, 81, 59–66.

G. S. BECKER, K. M. M., AND R. TAMURA. (1990): "Human capital, fertility, and economic growth," *Journal of Political Economy*, 98, S12–37.

H. LI, J. Z., AND Y. ZHU. (2008): "The quantity-quality trade-off of children in a developing country: Identification using chinese twins," *Demography*, 45, 223–43.

HANUSHEK, E. A. (1992): "The trade-off between child quantity and quality," *Journal of Political Economy*, 100, 84–117.

HAUSER, R. M., AND H. D. KUO. (1998): "Does the gender composition of sibships affect women's educational attainment?" *The Journal of Human Resources*, 33, 644–57.

IACOVOU, M. (2001): "Family composition and children's educational outcomes," *Institute for Social and Economic Research, ISER working papers,*.

KAESTNER, R. (1997): "Are brothers really better? Sibling sex composition and educational achievement revisited," *The Journal of Human Resources*, 32, 250–84.

MOGSTAD, M., AND M. WISWALL. (2016): "Testing the quantity–quality model of fertility: Estimation using unrestricted family size models," *Quantitative Economics*, 7, 157–92.

MORDUCH, J. (2000): "Sibling rivalry in africa," *American Economic Review*, 90, 405–9.

NANCY, Q. (2004): "Quantity-quality and the one child policy: The postive effect of family size on education in china,"

R. GUO, J. Y., AND J. ZHANG. (2017): "Family size, birth order, and tests of the quantity–quality model," *Journal of Comparative Economics*, 45, 219–24.

---. (2020): "The Child Quantity–Quality Trade-Off," in *Handbook of labor, human resources and population economics*, ed. by Zimmermann, K. F. Cham: Springer International Publishing, 1–23.

ROBERT, W. (1973): "A new approach to the economic theory of fertility behavior," *Journal of Political Economy*, 81, S14–64.

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, 1149–74.

S. E. BLACK, P. J. D., 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*, 120, 669–700.

---. (2010): "Small family, smart family? Family size and the IQ scores of young men," *The Journal of Human Resources*, 45, 33–58.

ZAJONC, R. B., AND G. B. MARKUS. (1975): "Birth order and intellectual development." *Psychological review*, 82, 74.

6 Appendix A

Table 3: Descriptive statistics by diploma status

label	levels	diploma	no_diploma	Total	p
age	Mean (SD)	24.1 (3.8)	24.0 (4.0)	24.0 (3.9)	0.497
birth_year	Mean (SD)	1985.9 (7.9)	1984.3 (7.5)	1984.9 (7.7)	<0.001
residence	city	788 (50.1)	1151 (44.6)	1939 (46.6)	<0.001
	village	352 (22.4)	1019 (39.5)	1371 (33.0)	
	1mln	155 (9.8)	173 (6.7)	328 (7.9)	
	MSK	223 (14.2)	176 (6.8)	399 (9.6)	
	SPB	56 (3.6)	64 (2.5)	120 (2.9)	
kids_num	1	689 (43.8)	889 (34.4)	1578 (38.0)	<0.001
	2	741 (47.1)	1194 (46.2)	1935 (46.5)	
	3	120 (7.6)	349 (13.5)	469 (11.3)	
	>=4	24 (1.5)	151 (5.8)	175 (4.2)	
kid_order	1	1139 (72.4)	1684 (65.2)	2823 (67.9)	<0.001
	2	393 (25.0)	707 (27.4)	1100 (26.5)	
	>=3	42 (2.7)	192 (7.4)	234 (5.6)	
nuclear_family	0	217 (13.8)	363 (14.1)	580 (14.0)	0.846
	1	1357 (86.2)	2220 (85.9)	3577 (86.0)	
parents_education	1 diploma	506 (32.1)	535 (20.7)	1041 (25.0)	<0.001
	2 diplomas	465 (29.5)	209 (8.1)	674 (16.2)	
	no diplomas	603 (38.3)	1839 (71.2)	2442 (58.7)	
lag_expenses	I Q	31 (3.9)	116 (8.8)	147 (6.9)	<0.001
	II Q	134 (16.7)	284 (21.5)	418 (19.7)	
	III Q	255 (31.7)	417 (31.6)	672 (31.7)	
	IV Q	384 (47.8)	501 (38.0)	885 (41.7)	
is_separated	0	1473 (93.6)	2356 (91.2)	3829 (92.1)	0.007
	1	101 (6.4)	227 (8.8)	328 (7.9)	
profession_type	not qualifed	132 (8.4)	813 (31.5)	945 (22.7)	<0.001
	qualified	892 (56.7)	858 (33.2)	1750 (42.1)	
	unemployed	550 (34.9)	912 (35.3)	1462 (35.2)	
gender	female	871 (55.3)	1071 (41.5)	1942 (46.7)	<0.001
	male	703 (44.7)	1512 (58.5)	2215 (53.3)	

Note: p-values of tests (t-test and chi-square test) for differences between graduates and non-graduates are reported.
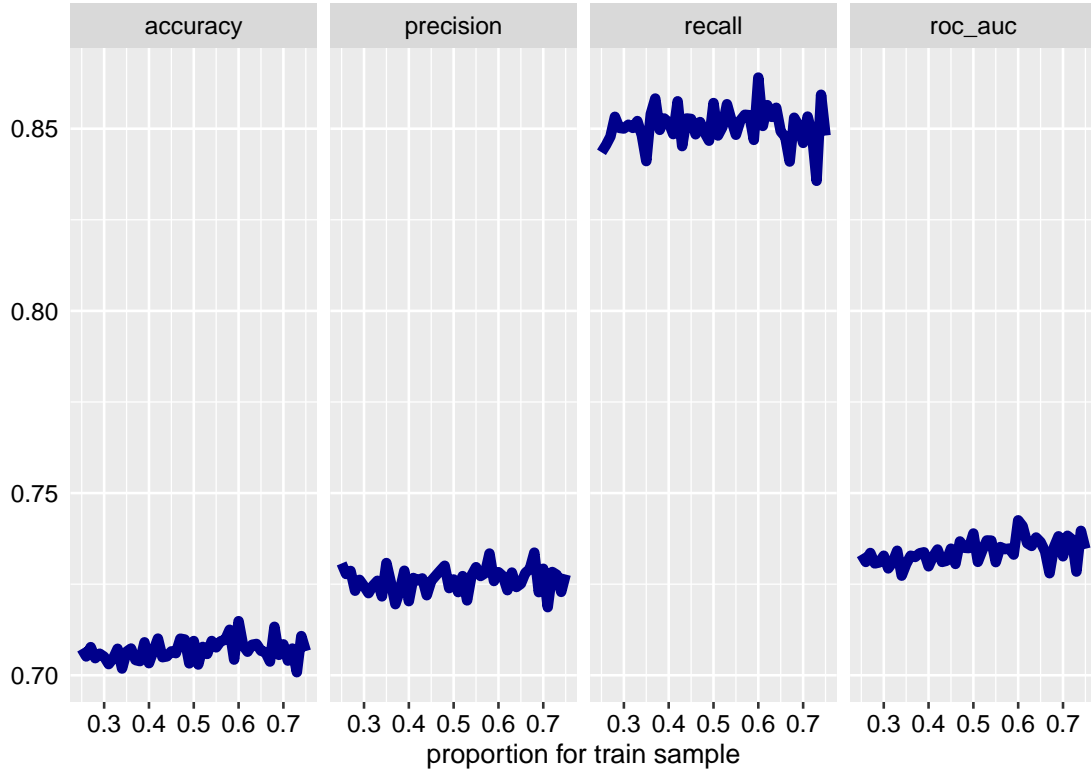
7 Appendix B

Figure 4: Full sample, quality metrics (threshold=0.5)

Table 4: estimates, full sample

term	kid_order2	kid_order>=3	kids_num2	kids_num3	kids_num>=4	nobs
+kids num	NA	NA	-0.235 (0.07)***	-0.844 (0.119)***	-1.6 (0.227)***	4157
+kid order	-0.044 (0.088)	-0.383 (0.218)	-0.215 (0.081)**	-0.72 (0.139)***	-1.389 (0.254)***	4157
+gender	-0.081 (0.089)	-0.412 (0.22)	-0.184 (0.082)*	-0.682 (0.14)***	-1.408 (0.257)***	4157
+birth year trunc	-0.081 (0.089)	-0.412 (0.22)	-0.184 (0.082)*	-0.682 (0.14)***	-1.408 (0.257)***	4157
+parents education	-0.068 (0.095)	-0.387 (0.23)	-0.222 (0.087)*	-0.633 (0.148)***	-1.193 (0.268)***	4157
+other gender siblings	-0.062 (0.096)	-0.378 (0.23)	-0.243 (0.096)*	-0.669 (0.162)***	-1.241 (0.282)***	4157
+is nuclear family	-0.063 (0.096)	-0.381 (0.231)	-0.241 (0.096)*	-0.664 (0.162)***	-1.238 (0.282)***	4157
+residence	-0.055 (0.096)	-0.348 (0.232)	-0.244 (0.097)*	-0.57 (0.164)***	-1.163 (0.287)***	4157
+lag expenses	0.081 (0.133)	-0.159 (0.291)	-0.343 (0.143)*	-0.838 (0.238)***	-1.228 (0.377)**	2122

Note: each row represents addition of new variable to those that have been added earlier.

Table 5: estimates, sample of not separated individuals

term	kid_order2	kid_order>=3	kids_num2	kids_num3	kids_num>=4	nobs
+kids num	NA	NA	-0.256 (0.073)***	-0.835 (0.124)***	-1.581 (0.233)***	3829
+kid order	-0.067 (0.092)	-0.368 (0.224)	-0.226 (0.084)**	-0.705 (0.145)***	-1.366 (0.263)***	3829
+gender	-0.102 (0.093)	-0.401 (0.226)	-0.196 (0.085)*	-0.674 (0.146)***	-1.388 (0.265)***	3829
+birth year trunc	-0.102 (0.093)	-0.401 (0.226)	-0.196 (0.085)*	-0.674 (0.146)***	-1.388 (0.265)***	3829
+parents education	-0.088 (0.099)	-0.359 (0.237)	-0.241 (0.091)**	-0.647 (0.154)***	-1.207 (0.277)***	3829
+other gender siblings	-0.081 (0.1)	-0.348 (0.237)	-0.266 (0.1)**	-0.69 (0.17)***	-1.264 (0.293)***	3829
+is nuclear family	-0.083 (0.1)	-0.353 (0.237)	-0.262 (0.1)**	-0.681 (0.17)***	-1.258 (0.294)***	3829
+residence	-0.075 (0.101)	-0.31 (0.239)	-0.265 (0.101)**	-0.582 (0.172)***	-1.193 (0.298)***	3829
+lag expenses	0.075 (0.142)	-0.071 (0.307)	-0.397 (0.154)**	-0.933 (0.257)***	-1.356 (0.402)***	1871

Note: each row represents addition of new variable to those that have been added earlier.

Figure 5: profiling CI for the full sample, all controls except lag expenses

univ_diploma: OR (95% CI, p–value)

|		
father_univ_diploma	1	2.37 (2.01–2.80, p<0.001)
mother_univ_diploma	1	2.63 (2.23–3.10, p<0.001)
gender	male	0.51 (0.44–0.59, p<0.001)
residence	city	–
	village	0.62 (0.52–0.74, p<0.001)
	1mln	1.20 (0.92–1.57, p=0.185)
	MSK	1.55 (1.21–1.99, p<0.001)
	SPB	0.73 (0.48–1.10, p=0.137)
birth_year_trunc	1970	–
	1975	1.54 (1.15–2.06, p=0.003)
	1980	1.69 (1.29–2.23, p<0.001)
	1985	1.75 (1.33–2.29, p<0.001)
	1990	1.48 (1.11–1.98, p=0.007)
	1995	3.13 (2.30–4.28, p<0.001)
	2000	5.06 (2.91–8.94, p<0.001)
kids_num	1	–
	2	0.77 (0.63–0.93, p=0.009)
	3	0.56 (0.40–0.78, p=0.001)
	>=4	0.30 (0.17–0.54, p<0.001)
kid_order	1	–
	2	0.93 (0.76–1.13, p=0.457)
	>=3	0.73 (0.46–1.17, p=0.195)
nuclear_family	1	1.10 (0.90–1.36, p=0.357)
other_gender_siblings	1	1.12 (0.92–1.37, p=0.270)

Odds ratio (95% CI, log scale)
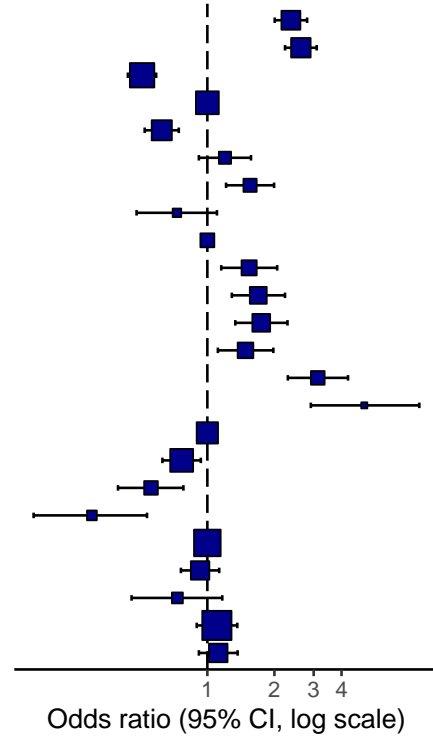
Figure 6: profiling CI for the sample of not separated individuals, all controls except lag expenses

8 Appendix C: cohort effects

Table 6: estimates, full sample of individuals born in 1970-1979

term	kid_order2	kid_order>=3	kids_num2	kids_num3	kids_num>=4	nobs
+kids num	NA	NA	-0.191 (0.137)	-0.564 (0.253)*	-1.282 (0.452)**	1105
+kid order	-0.088 (0.189)	-1.861 (1.056)	-0.162 (0.15)	-0.408 (0.269)	-0.955 (0.474)*	1105
+gender	-0.139 (0.19)	-1.823 (1.059)	-0.128 (0.152)	-0.347 (0.271)	-0.905 (0.476)	1105
+birth year trunc	-0.139 (0.19)	-1.823 (1.059)	-0.128 (0.152)	-0.347 (0.271)	-0.905 (0.476)	1105
+parents education	-0.13 (0.209)	-1.662 (1.105)	-0.262 (0.17)	-0.295 (0.294)	-0.7 (0.5)	1105
+other gender siblings	-0.135 (0.209)	-1.688 (1.111)	-0.2 (0.188)	-0.189 (0.325)	-0.58 (0.523)	1105
+is nuclear family	-0.134 (0.209)	-1.687 (1.109)	-0.206 (0.188)	-0.201 (0.326)	-0.597 (0.525)	1105
+residence	-0.123 (0.21)	-1.671 (1.116)	-0.223 (0.189)	-0.174 (0.328)	-0.543 (0.534)	1105
+lag expenses	-0.245 (0.384)	-1.048 (1.371)	-0.296 (0.41)	-0.163 (0.611)	-1.308 (1.196)	269

Note: each row represents addition of new variable to those that have been added earlier.

Table 7: estimates, full sample of individuals born in 1980-1989

term	kid_order2	kid_order>=3	kids_num2	kids_num3	kids_num>=4	nobs
+kids num	NA	NA	-0.156 (0.107)	-0.767 (0.17)***	-1.468 (0.311)***	1845
+kid order	-0.095 (0.127)	-0.211 (0.291)	-0.11 (0.124)	-0.663 (0.205)**	-1.315 (0.362)***	1845
+gender	-0.13 (0.128)	-0.28 (0.293)	-0.066 (0.125)	-0.611 (0.207)**	-1.29 (0.363)***	1845
+birth year trunc	-0.13 (0.128)	-0.28 (0.293)	-0.066 (0.125)	-0.611 (0.207)**	-1.29 (0.363)***	1845
+parents education	-0.132 (0.135)	-0.281 (0.308)	-0.008 (0.133)	-0.553 (0.217)*	-1.058 (0.379)**	1845
+other gender siblings	-0.088 (0.137)	-0.229 (0.31)	-0.127 (0.145)	-0.752 (0.239)**	-1.337 (0.405)***	1845
+is nuclear family	-0.086 (0.137)	-0.227 (0.31)	-0.129 (0.145)	-0.755 (0.24)**	-1.341 (0.405)***	1845
+residence	-0.098 (0.138)	-0.195 (0.312)	-0.148 (0.147)	-0.683 (0.242)**	-1.337 (0.411)**	1845
+lag expenses	0.03 (0.196)	-0.1 (0.411)	-0.107 (0.224)	-0.841 (0.362)*	-1.054 (0.524)*	935

Note: each row represents addition of new variable to those that have been added earlier.

Table 8: estimates, full sample of individuals born in 1990 or later

term	kid_order2	kid_order>=3	kids_num2	kids_num3	kids_num>=4	nobs
+kids num	NA	NA	-0.385 (0.128)**	-1.16 (0.224)***	-2.128 (0.489)***	1207
+kid order	0.095 (0.165)	-0.171 (0.403)	-0.434 (0.154)**	-1.132 (0.275)***	-2.047 (0.549)***	1207
+gender	0.079 (0.168)	-0.089 (0.407)	-0.443 (0.157)**	-1.162 (0.278)***	-2.274 (0.559)***	1207
+birth year trunc	0.079 (0.168)	-0.089 (0.407)	-0.443 (0.157)**	-1.162 (0.278)***	-2.274 (0.559)***	1207
+parents education	0.105 (0.177)	-0.018 (0.42)	-0.546 (0.165)***	-1.098 (0.288)***	-2.128 (0.584)***	1207
+other gender siblings	0.087 (0.178)	-0.049 (0.423)	-0.486 (0.179)**	-0.986 (0.315)**	-1.982 (0.608)**	1207
+is nuclear family	0.09 (0.178)	-0.076 (0.424)	-0.487 (0.179)**	-0.959 (0.316)**	-2.011 (0.608)***	1207
+residence	0.11 (0.18)	-0.064 (0.426)	-0.43 (0.182)*	-0.785 (0.321)*	-1.83 (0.619)**	1207
+lag expenses	0.294 (0.213)	-0.067 (0.474)	-0.611 (0.219)**	-1.006 (0.385)**	-1.707 (0.679)*	918

Note: each row represents addition of new variable to those that have been added earlier.