Workforce quality in early years interventions: Evidence from a large-scale home visiting program

Sarah Cattan, Gabriella Conti and Christine Farquharson

June 4, 2023

Abstract

Targeted early years programmes can have significant benefits for children's development; however, there are many examples of early interventions that have failed to live up to their promise, particularly when delivered at scale. Understanding the inputs into a successful early years programme is therefore essential to reform existing interventions and guide investment into the most promising programmes. In this paper, we consider the role of a crucial input: workforce effectiveness. We evaluate the degree of heterogeneity in workforce effectiveness in the context of the highly trained workforce employed by a successful, scaled home visiting programme in England. Using the quasi-random assignment of workers to families for identification, we estimate each worker's value-added in promoting children's physical, cognitive and socio-emotional development. We find evidence of substantial heterogeneity in workforce effectiveness; for example, a one-standard deviation increase in effectiveness leads to a 0.24SD increase in cognitive performance at age 2, and a 0.29SD increase in socio-emotional development. However, despite having access to unusually detailed data on worker characteristics and on process quality, we are only able to explain around 15% of the variation in worker effectiveness. Overall, our results show that there is substantial heterogeneity in effectiveness even among highly skilled workers, but we are only starting to understand its determinants.

*Cattan: Institute for Fiscal Studies and IZA; Conti: University College London and Institute for Fiscal Studies. Farquharson: Institute for Fiscal Studies. The authors gratefully acknowledge funding from the Tavistock and Portman NHS Foundation Trust, the ESRC Centre for the Microeconomic analysis of Public Policies at the Institute for Fiscal Studies, the British Academy through Cattan's Postdoctoral Fellowship, and the NORFACE DIAL for the project "Growing up Unequal? The Origins, Dynamics and Lifecycle Consequences of Childhood Inequalities". We are indebted to Ruth Rothman, Phil Howlin and Andreea Moise from the Family Nurse Partnership National Unit for answering our numerous questions about the data and the programme. We thank seminar participants at the "Medical Interventions and Socioeconomic Outcomes" Workshop, at the UCL 2017 Research Day Workshop, and at the University of Copenhagen, Venice and Queen Mary for useful comments.

1 Introduction

Children's experiences in their early years strongly influence their later outcomes across a range of areas. Children born in more vulnerable families risk failing to reach their developmental potential, in part because they are exposed to unstable, unsafe and non-stimulating environments during some of their most formative years (Almond and Currie, 2011; Cunha et al., 2006). These factors are, for the child, purely a matter of chance. Yet, they have profound consequences for their lifelong well-being.

Policymakers around the world are grappling with how to deal with this unfair lottery of birth and thus foster intergenerational mobility while improving outcomes for the most disadvantaged. One of the most promising tools is home visiting programmes, which are backed by a large body of international, interdisciplinary evidence (Avellar and Supplee, 2013; Peacock et al., 2013). However, there are well-known challenges to scaling up such programmes, and evaluations of interventions operating at full scale frequently find weaker and more heterogeneous impacts than in earlier, smaller trial settings.

Understanding the role of different programme inputs is therefore essential both to modifying existing interventions and to choosing new programmes that are most likely to succeed. A parallel literature on schools has grappled with this same question. This research highlights the vital role of teachers in explaining pupil achievement as well as the large amount of heterogeneity in teacher effectiveness (Hanushek and Rivkin, 2006, 2012).

In this paper, we therefore focus on quantifying the extent and importance of heterogeneity in workforce effectiveness in early years programmes. Our setting is the Family Nurse Partnership (FNP) programme in England. This is a large-scale, structured home visiting programme for first-time teenage mothers, based on the successful Nurse Family Partnership in the U.S. (Olds et al., 1999). The programme employs 'family nurses' to deliver roughly two home visits a month to each mother, lasting from enrollment (during pregnancy) through to the child's second birthday.

Using unique high-quality data on all the family nurses and the clients in the ten years since the beginning of the programme, we adapt the research design developed in the teacher effectiveness literature and apply it for the first time in the context of an early intervention. For identification, we exploit the quasi-random assignment of family nurses to mothers (conditional on a small number of known and observed assignment rules).

We show that there is substantial heterogeneity in the impact of family nurses on children's health, cognitive and socio-emotional development. However, despite access to exceptionally detailed information about the family nurses and their on-the-job performance, we are nevertheless only able to explain a small amount of this variation.

We proceed in two stages. First, we estimate the overall impact of family nurses on children's health at birth, and on their cognitive and socio-emotional development at ages 1 and 2. To measure health at birth, we construct a factor combining birth weight, gestational age and time spent in neonatal intensive care. Our cognitive and socio-emotional measures are derived from the Ages and Stages Questionnaires, a validated instrument for use in the early years. We then use exceptionally rich data on family nurses - including their objective characteristics, their supervisors' evaluations of them, and their own reports after each visit - to attempt to identify predictors of more effective family nurses.

Identifying the effectiveness of the workforce delivering early interventions is challenging for the same reasons that it is in the teacher literature. Specifically, caregivers or family nurses are seldom randomly assigned to children or their mothers. This makes it difficult to separate their effects from that of the child's unobserved endowments and of other inputs that are relevant to explain children's outcomes. To overcome this problem in the teacher quality literature, studies either exploit situations in which teachers are (voluntarily or involuntarily) randomly assigned to students (e.g. Kane et al. (2008); Deming (2014); Angrist et al. (2017); Araujo et al. (2016)) or use one or several lagged outcomes to proxy for the child's unobserved endowments and the history of unobserved inputs (Todd and Wolpin, 2003). This "value-added" approach has been implemented in a variety of settings to estimate the effect of teachers on academic gains (see Koedel et al. (2015) for a review) and more recently on adult outcomes (Chetty et al., 2014a,b).

In our paper, we instead exploit a unique feature of the assignment process of the family nurses to the mothers in the FNP to identify the nurses' impacts on children's outcomes. The assignment process occurs after a mother is referred to her local FNP team, but before she completes detailed intake paperwork. This means that team leaders (supervisors) have access only to basic demographic information about each mother. The supervisor assigns a nurse to a new mother primarily based on the complexity of the nurse's existing caseload relative to that of other family nurses in her team. Conditional on this, as well as the month of enrollment and basic mother demographics, the choice of the family nurse assigned should be as good as random. We use rich data on the characteristics of family nurses and of mothers to construct these measures of caseload complexity at each month of the programme. We present evidence strongly suggesting that, conditional on a family nurse's relative caseload complexity at the time of a new mother's enrollment, there is no significant correlation between the characteristics of the mother and the characteristics of her assigned family nurse.

Using this identification strategy, we show a moderate degree of heterogeneity in family nurses' ability to promote better health at birth: a one-standard deviation (SD) increase in family nurse quality leads to a 0.12 SD increase in health at birth (which we construct from information on birthweight, gestational age and time spent in neonatal intensive care). This is about 20% more heterogeneity than would be expected to arise if there were no link between the family nurse and the child's birth outcomes. We find a much larger degree of heterogeneity when considering cognitive and socio-emotional development at ages 1 and 2. For example, a one-standard deviation increase in family nurse effectiveness raises the child's age 2 cognitive performance by 0.24SD, and her age 2 socio-emotional development by 0.29SD. For these later outcomes, the degree of heterogeneity between workers is around twice what would be expected to arise by chance.

We then turn our attention to understanding the predictors of family nurse quality.

Studies in developmental psychology and in economics increasingly conclude that so-called "structural indicators", such as caregivers' qualifications, experience and group size, are less predictive of children's outcomes than the quality of interactions between caregivers and children. This literature, however, is far from conclusive and, with a few exceptions, only focuses on teachers in pre-school and kindergarten settings (Araujo et al., 2016; Blanden et al., 2017). Very little is known about the best predictors of effectiveness in the early years workforce, especially in home-visiting interventions targeting younger children. We fill this gap by exploiting exceptionally rich administrative data to explore the predictive power of family nurses' objective characteristics (e.g. age, ethnicity, experience and training); "process quality" data collected from their supervisor's evaluation, conducted by observing one or more home visit; and monitoring data on the content of the home visits reported by the family nurse herself (e.g. duration of visits and ratings of how engaged the mother was in the visit). Despite the richness of our data, we are only able to predict a modest degree of the variation in family nurses' effectiveness.

We show that we can explain less than 15% of this variation in family nurse effectiveness, despite having a very rich set of observable nurse characteristics; supervisor assessments made during accompanied visits; and detailed information on programme implementation. We find that it is this last set of predictors which is most closely related to estimated nurse quality - perhaps surprisingly, assessments made by supervisors bear almost no relation to the quality measures we estimate.

The rest of the paper is organized as follows. Section section 2 provides some background on the FNP, while section 3 describes the data. Section section 4 presents our conceptual framework, formalizes the parameters that we aim to identify, and presents supportive evidence for our identification strategy. Section section 5 presents our estimates of family nurse effectiveness for children's outcomes. Section 6 discusses the results of our analysis of the determinants of family nurse quality. Section 7 concludes.

2 The Family Nurse Partnership programme

2.1 History of the FNP in England

The Family Nurse Partnership (FNP) is a large-scale home visiting programme, focussing on encouraging behaviour change among a highly disadvantaged client base. The programme has three main goals: to improve the outcomes of pregnancy by helping women improve their prenatal health; to improve children's health and development by enabling parents to provide more competent care for their children; and to improve women's life course by supporting them to plan any subsequent pregnancies, finish their education and/or find employment. To achieve these aims, it offers first-time teenage mothers regular home visits with highly-trained family nurses covering a structured curriculum.

The programme is modelled on the successful Nurse Family Partnership in the U.S., which has been shown (via a randomised controlled trial) to have consistent and enduring benefits for maternal and child health (Olds et al., 1999). RCT evaluation of the FNP in England found no significant impact on the trial's pre-designated primary outcomes (birth weight, smoking, repeated pregnancy and hospital admissions) (Olds, 2015). However, these trials did show substantial benefits for child cognitive development (which was considered a secondary outcome) (Robling et al., 2015).[1] In this paper, we therefore focus on analysing the role of workforce quality in promoting child outcomes.

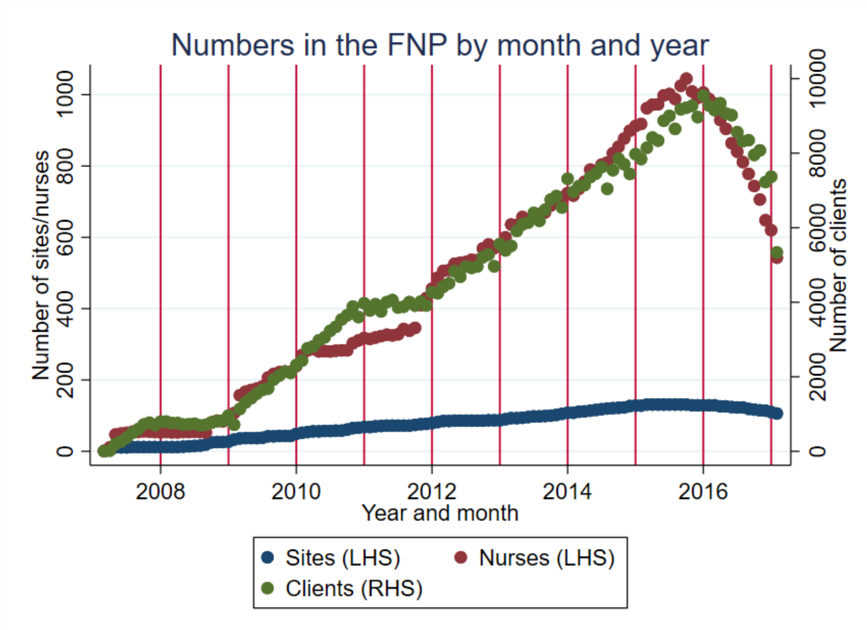
The FNP programme in England operates at significant scale. In the decade following its introduction as a pilot in 2007, the programme served over 32,000 young mothers (clients) across 131 sites.[2] Figure 1 shows the number of active mothers, family nurses and sites for each year and month between February 2007 and February 2017. After its initial trial period in 2007 and 2008, the programme rolled out quickly. At its peak in 2016, the FNP

[1] The FNP trial in England is also notable for the high number of visits delivered to the control group through "usual care", potentially attenuating the measured effectiveness of the programme for the treatment group.

[2] In 2010, the programme was also extended to cover Scotland and Northern Ireland. However, our data are limited to the English programme.

had around 10,000 active clients and 1,000 nurses across 131 sites.

Figure 1: Number of nurses, clients and sites over time



Note: The figure shows the distribution. Source: Authors' calculations using data provided by the Department for Health and Social Care.

2.2 An overview of the FNP programme

While there is a wide range of other early years programmes in operation in England, the FNP stands out as one of the most targeted, intensive, structured and scaled-up programmes on offer. The programme serves highly disadvantaged clients: first-time teenage mothers.

To be eligible for the FNP, mothers must have been aged 19 or under at last menstrual period; have a first pregnancy confirmed by health services (those whose previous pregnancies ended in miscarriage, termination or stillbirth are also eligible); be less than 28 weeks of gestation; live within the catchment area covered by the local FNP team; and have no planned adoption at enrolment. Non-English speaking clients and clients with learning disabilities are also eligible. Table 1 highlights the high levels of social, economic and health disadvantage among FNP clients. For example, 25% of mothers in our sample are aged 16 or younger at enrolment. Despite this, over half of mothers are not in education, employment or training. While over three quarters of mothers have a current partner, almost a fifth of mothers are not in weekly contact with their child's biological father.

The FNP programme offers these young mothers a very structured, intensive intervention of regular home visits by a trained family nurse (FN). Mothers enrol during pregnancy (no later than 28 weeks) and receive up to 64 visits, continuing through their child's second birthday. The visits incorporate content from the five strands of the programme: improving mothers' physical and mental health; bolstering mothers' knowledge of child development and confidence in parenting; promoting a healthy home environment for children; supporting mothers to set and meet their own life goals (e.g. in family planning, education or employment); and encouraging mothers to nurture and access their own support networks.

As Figure 2 shows, the emphasis of the FNP curriculum shifts somewhat depending on the age of the child: before birth, there is greater focus on maternal physical health (e.g. smoking cessation or support with diet). Once the child is born, the emphasis shifts and lessons on the maternal role and child development take up around 40% of the visiting time. While the FNP curriculum sets out the intended content for each visit, family nurses are encouraged to use their own professional judgement to customise the programme so it meets the strengths and challenges of each family.

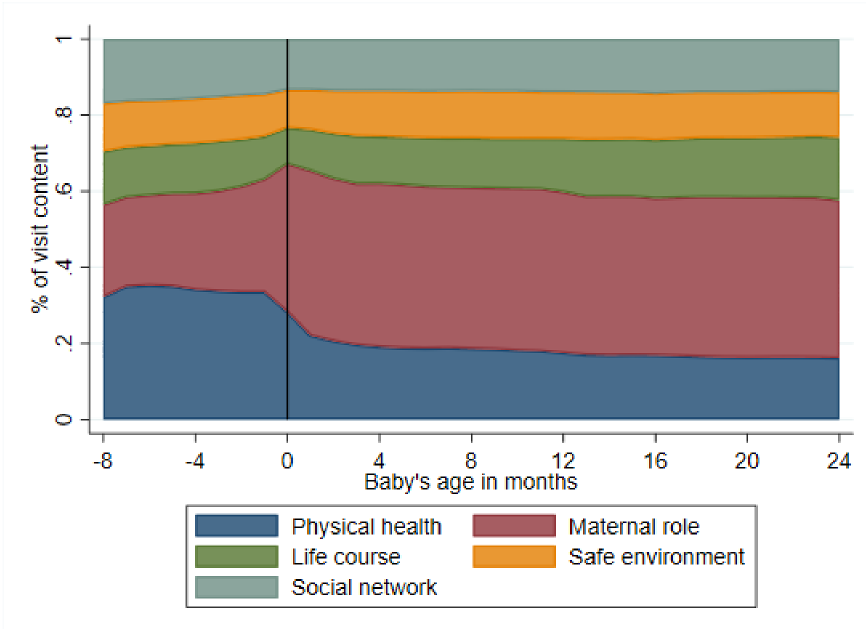
The median client receives two visits per month, with 20% of clients receiving three or more visits. As Figure 3 shows, visits are on average slightly longer in the months before birth, lasting just under 80 minutes; after birth, the average visit lasts just over 70 minutes. Nurses are able to cover most of the planned content in that time; on average, nurses report getting through just under 95% of the content they had planned to cover (Figure 3).

The FNP workforce The FNP visits are delivered by family nurses (FNs). These are highly qualified professionals who are required to have nursing or midwifery qualifications; to be registered with the Nursing and Midwifery Council; and to be educated to degree

Table 1: Summary statistics: Mothers' characteristics at FNP enrolment

	Full sample		Birth outcomes sample		Age 2 cognitive sample	
Characteristic	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Age (months)	215.2	16.2	214.4	16.3	214.1	15.9
Gestational age (weeks)	17.9	4.9	18.0	4.9	17.8	4.8
Primary language English	0.951	0.216	0.958	0.201	0.965	0.184
Ethnicity: Black	0.054	0.225	0.054	0.226	0.051	0.220
Ethnicity: Asian	0.026	0.158	0.025	0.157	0.024	0.153
Ethnicity: Mixed/Other	0.070	0.255	0.066	0.248	0.061	0.239
Mother has current partner	0.769	0.422	0.766	0.423	0.777	0.416
Contact with bio dad: Never	0.139	0.346	0.138	0.345	0.129	0.335
Contact with bio dad: ¡ Weekly	0.057	0.233	0.058	0.235	0.054	0.227
Contact with bio dad: Weekly	0.104	0.306	0.108	0.311	0.104	0.305
Contact with bio dad: Daily	0.477	0.499	0.480	0.500	0.498	0.500
Residence: Private renter	0.212	0.409	0.209	0.407	0.207	0.405
Residence: Social renter	0.459	0.498	0.462	0.499	0.470	0.499
Residence: Non-response	0.144	0.351	0.139	0.346	0.120	0.325
Residence: Non-traditional	0.082	0.275	0.070	0.255	0.054	0.226
Household: Own partner	0.286	0.452	0.279	0.449	0.283	0.451
Household: Own mother	0.541	0.498	0.552	0.497	0.567	0.495
Not in educ/employment/training	0.527	0.499	0.527	0.499	0.504	0.500
Lacks good qualifications	0.681	0.466	0.689	0.463	0.679	0.467
Any vocational qualification	0.435	0.496	0.427	0.495	0.439	0.496
Ever in paid work	0.461	0.499	0.442	0.497	0.455	0.498
Currently in paid work	0.181	0.385	0.169	0.375	0.178	0.382
Income only from benefits	0.322	0.467	0.329	0.470	0.321	0.467
Any social care	0.171	0.377	0.157	0.364	0.123	0.329
Any social services	0.492	0.500	0.506	0.500	0.497	0.500

Figure 2: FNP programme content, by age of child



Note: The figure shows the proportion of time each visit allocated to each of five strands of programme content, averaged across all visits. Source: Authors' calculations using data provided by the Department for Health and Social Care.

level. In addition to these requirements, FNs undergo extensive additional training to be part of the programme. Reflecting their high levels of training and experience, FNs are on a higher pay band than midwives, staff nurses or most health visitors in the English National Health System. As Table 2 sets out, around half of family nurses are qualified as a general nurse; just over a third are qualified midwives, and half hold a degree in health visiting. In accordance with the FNP's target population, over 70% of family nurses had experience working with teenage parents before joining the programme.

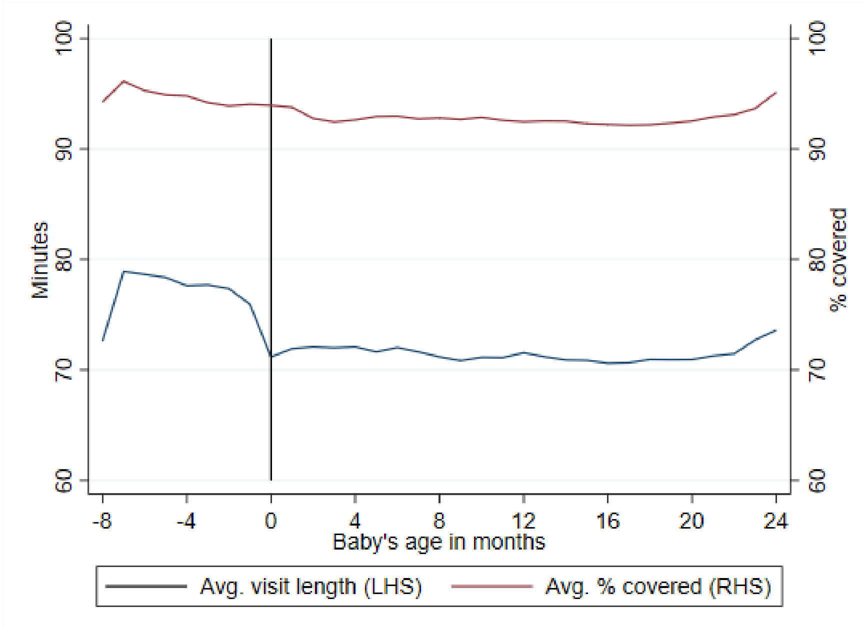
In each site, the FNs carry a maximum caseload of 30 clients. However, as Figure 4 shows, such high caseloads are unusual; the median nurse has 14 clients at any given time and carries out 24 visits per month (just under two per client). They work in teams led by a Family Nurse Supervisor, typically with up to eight FNs and one SV.[3] In addition to carrying a small caseload of their own clients, supervisors are responsible for the management and

[3]There can be multiple teams per site, depending on local demand, but data on the assignment of FNs to individual supervisors is patchy.

Table 2: Summary statistics: Family nurses' characteristics at time of joining FNP

	Full sample		Birth outcomes sample		Age 2 cognitive sample	
Characteristic	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Supervisor	0.127	0.333	0.104	0.306	0.093	0.290
Female	0.989	0.105	0.991	0.095	0.993	0.085
Non-white	0.083	0.276	0.073	0.260	0.069	0.253
Age 40+	0.695	0.460	0.758	0.429	0.753	0.432
Age 50+	0.243	0.429	0.262	0.441	0.260	0.439
Qual: General nurse	0.498	0.500	0.557	0.497	0.543	0.499
Qual: Adult nurse	0.245	0.430	0.236	0.425	0.214	0.411
Qual: Mental health nurse	0.056	0.231	0.047	0.211	0.049	0.217
Qual: Children's nurse	0.217	0.412	0.210	0.408	0.230	0.422
Qual: Learning disability	0.009	0.093	0.009	0.093	0.007	0.081
Qual: Midwife	0.354	0.478	0.359	0.480	0.349	0.477
Qual: School nurse	0.048	0.215	0.047	0.211	0.046	0.210
Qual: Health visitor	0.150	0.357	0.157	0.365	0.151	0.359
Educ: BSc school nursing	0.088	0.284	0.087	0.283	0.089	0.285
Educ: BSc health visiting	0.565	0.496	0.554	0.498	0.543	0.499
Educ: MSc credits	0.208	0.406	0.201	0.401	0.207	0.406
Educ: MSc degree	0.098	0.297	0.093	0.291	0.082	0.275
Experience: Community work	0.954	0.210	0.971	0.168	0.974	0.160
Experience: Teen parents	0.704	0.457	0.726	0.447	0.730	0.445
Experience: Safeguarding	0.035	0.184	0.032	0.176	0.033	0.179
Experience: CAMHS	0.107	0.309	0.128	0.335	0.122	0.327
Training: Safeguarding	0.462	0.499	0.513	0.501	0.533	0.500
Training: Safeguarding supervisor	0.179	0.384	0.172	0.378	0.135	0.342
Training: Common Assessment	0.556	0.497	0.653	0.477	0.684	0.466
Training: Management	0.092	0.289	0.073	0.260	0.066	0.248
Training: Other additional	0.155	0.362	0.201	0.401	0.234	0.424

Figure 3: Visit duration and coverage of planned content, by age of child



Note: The figure shows the average length of a visit (in minutes) and the average proportion of the planned content covered, by the age of the client's child. Both outcomes are recorded by the family nurse for each visit she conducts. Source: Authors' calculations using data provided by the Department for Health and Social Care.

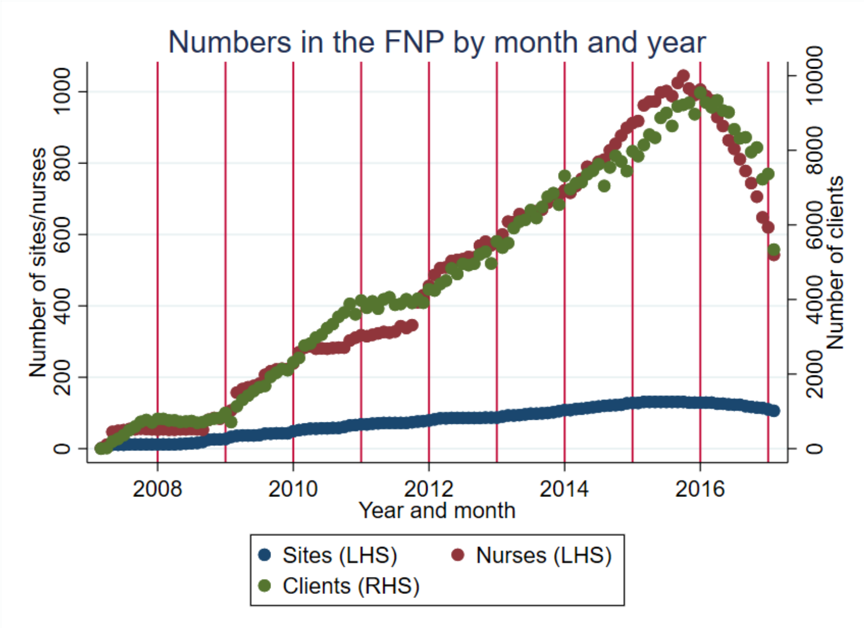
support of the FNs on their team. This includes regularly attending home visits with each FN (around twice a year) to provide an assessment of the FN's performance and give the FN detailed feedback. Supervisors are also responsible for allocating FNs to new clients, as we describe in greater detail now.

2.3 Enrollment in the FNP and nurse-client assignment

Our empirical strategy relies on the quasi-random assignment of clients to nurses (conditional on a limited set of observables). In this section we therefore provide further detail about the process through which nurses are assigned to mothers, before empirically testing the plausibility of quasi-random assignment in section 4.

Mothers joining the FNP go through four stages. In the initial *referral* stage, the mother

Figure 4: Number of nurses, clients and sites over time



Note: The figure shows the total number of active sites, nurses and clients in each month. Source: Authors' calculations using data provided by the Department for Health and Social Care.

is referred to the programme by her midwife or doctor.[4] She fills out a referral form containing basic information: her contact details, hospital, date of birth, expected due date, language and religion, and residential status (see Appendix A for a sample referral form).

The mother is then assessed against the strict FNP eligibility criteria: mothers must be no more than 28 weeks pregnant and must be aged 19 or younger with no previous children. Assuming the mother meets these criteria, she is then issued an *invitation* to join the FNP. If she accepts, she then proceeds to the *allocation* stage. The supervisor at her local FNP site allocates her to one of the family nurses, based on the caseload of each of the FNs on the team and on a limited set of the mother's characteristics (including her age, expected due date, and language). Specifically, supervisors follow a three-step process: they identify nurses on their team who are not carrying a full caseload of clients; analyse the complexity of each nurse's caseload; and consider the geographic location, language and expected due date of the mother to determine which nurse is best able to take her on as a client.

[4]Mothers can also self-refer into the programme, but this is rare.

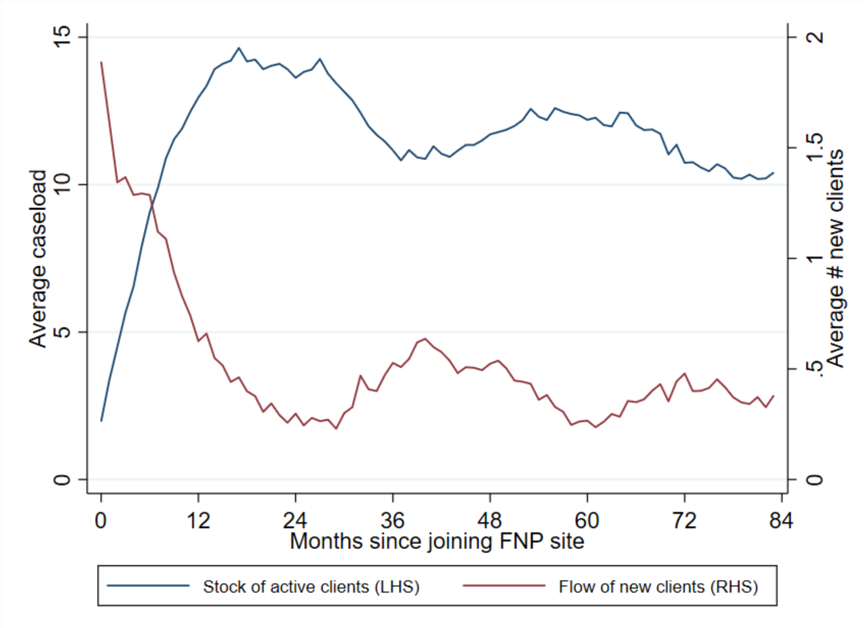
Finally, when the allocation of client to nurse has been made, the client and her family nurse together complete the *intake* paperwork. At this point, the nurse collects extensive additional information about her new client, including socio-economic demographics, health status and risk factors (such as smoking, drinking or domestic violence).

The sequencing of this four-stage process means that the decision of which nurse to allocate to a new client is taken at a point when supervisors have only a very limited set of information about the mother. In practice, the driving force behind the allocation decision is the caseload of the other nurses on the team. As Figure 5 shows (in blue), the average nurse takes just over a year to build up to a full caseload of around 14 mothers. Her caseload then plateaus for about 18 months before falling (as some of the mothers she works with complete the FNP programme). There remains some cyclicality in caseloads after this, but it is much less pronounced than the initial build-up period. The figure also shows the average number of new mothers a family nurse takes on each month (in red); this is inversely related to the nurse's caseload, showing that supervisors are much more likely to assign new mothers to nurses carrying smaller caseloads.

Stability of nurse-mother assignment In the majority of cases, mothers remain with their assigned nurse throughout their time in the FNP. Around 20% of mothers ever switch nurse; virtually none switch nurse multiple times. Just under half of the mothers switching nurse (44%) do so because their original family nurse stops working with the FNP at that site, either temporarily (for example, because of maternity leave or long-term illness) or permanently (due to leaving the FNP or switching site). In all our analysis, we focus on the family nurse initially assigned to the mother, since there is no guarantee of quasi-random assignment for subsequent nurses (when much more information about the client is available).

While mothers largely remain within the same FNP site that they enrol with, switching sites is somewhat more common among nurses. Leaving aside very infrequent 'filling in' for colleagues at other sites, 18% of nurses worked at multiple sites throughout their FNP career.

Figure 5: Family nurse caseloads by tenure in the FNP



Note: The figure shows the average caseload (number of active clients) and the average number of new clients assigned to a nurse, by the number of months since she joined the site. Source: Authors' calculations using data provided by the Department for Health and Social Care.

However, in the vast majority of cases, a nurse working at a secondary site did so to follow a specific mother. While the median nurse saw 21 mothers during her career at her main site, she saw only one mother at a secondary site. Empirically, as we will see in section 4, this means that we will not be able to separately identify the effectiveness of an individual nurse from the impact of her wider team. Rather, we focus on the total effectiveness of each family nurse - her own quality, plus any direct impact that the team has on individual families.

3 Data

In this paper, we use detailed anonymised administrative data from the FNP Information System (FNP IS). These data contain rich information on each client; on her child; on each of the family nurses; and on each visit (including which family nurse carried out the visit).[5]

[5]While the database also contains identifiers for supervisors, the data quality is much poorer; even after significant imputation, we are only able to confidently identify a family nurse's supervisor in 70% of nurse-

In this section, we provide an overview of the data we hold for each group.

While we have data available for the first decade of the programme (February 2007-February 2017), we restrict our analysis to clients and nurses who joined the FNP before 2014. This sample restriction ensures that we have enough time to observe children through to their second birthday. It also means that we focus on the period when the FNP programme was more consistently implemented; following the publication of the trial evaluation results in 2015, different FNP sites have adopted different approaches.

3.1 Data on children

One distinctive feature of the FNP IS data is the rich set of measures it contains on children's health at birth and their cognitive and non-cognitive development. Motivated by the results of the *Building Blocks* evaluation of the FNP, which found that the programme significantly improved children's development, we focus on exploring the role of family nurse quality in shifting child development in each of these domains.

Specifically, we construct five outcome variables, measured at different points in time. Our first measure of child development relates to their health stock at birth. To reduce the potential influence of measurement error, we use factor analysis to aggregate the 'signal' of health stock from three different birth outcomes: gestational age, birth weight, and the number of weeks spent in neonatal intensive care.

We complement this early health measure with four measures of children's later development: two cognitive, and two socio-emotional. Our cognitive measures are based on the Ages and Stages Questionnaire-3 (ASQ-3), which is collected at roughly 12 and 24 months.[6] The ASQ-3 is a commonly-used screening tool measuring child cognitive development on five

month cells.

[6] In practice, there is considerable variation in precisely when family nurses conducted the ASQ-3 assessments. Nurses were instructed to use the version of the ASQ survey most appropriate to the child's actual age. For our 12-month measure, we therefore pool the results for children assessed using the 10-month and the 14-month questionnaire. Where results are available for both questionnaires, we take a simple average of the two (residualised) scores. We conduct a similar procedure for the 24-month measure, in that case pooling results from the 20-month and 24-month questionnaires.

different domains: communication, gross motor skills, fine motor skills, problem solving, and personal social skills.

Our non-cognitive measures are based on the Ages and Stages Questionnaire-Social Emotional supplement (ASQ-SE), which is conducted at 12 and at 24 months. Its score also ranges from 0 to 60, where a higher score indicates worse socio-emotional development.[7]

Both the ASQ-3 and ASQ-SE have been widely analysed and both are reported to have high concurrent and predictive validity. For example, Salinas and Armijo (2017) shows that the ASQ-3 amongst children at 8, 18 and 30 months of age is highly predictive of their intelligence between 6 and 9 years of age, as measured by the Wechsler Intelligence Scale for Children-third edition (WISC-III). As measures of very early development, the ASQ questionnaires are normed to quite a narrow age range (4 months in the case of the ASQ-3). In addition, both scales are developmental screeners; this means they are primarily aimed at identifying children with developmental problems (and they are less good at discriminating among very advanced children). This results in a truncated distribution and a long left tail. We therefore residualise the ASQ and ASQ-SE measures on the child's sex and age in months at the time of test. We then winsorise the distribution at the 0.5th percentile of the distribution to avoid our regression estimates placing outsize importance on the children in the long left tail.[8]

3.2 Data on family nurses and on mothers

In addition to the rich data on children's outcomes, the FNP IS also contains detailed information about the characteristics of family nurses and their clients. We measure these

[7]We have reversed the scoring of the ASQ in our analysis for ease of interpretation so that a higher score indicates better development.

[8]Rather than using the full distribution of the ASQ and ASQ-SE, many previous studies have instead constructed categorical variables indicating whether the child is in the 'referral zone' for possible developmental problems. These referral zones are identified using cut-offs which are specific to the US population; no such value exists for the UK context. Further, binary outcomes are not easily used when estimating a large number of fixed effects (our measures of nurse effectiveness). We therefore follow a large literature, including Barnard et al. (2013) and Doyle et al. (2013), in using the full domain of ASQ scores and modelling them as continuous variables.

observable characteristics at the time a nurse enters the FNP (to avoid capturing any changes in training or experience as a result of participating in the programme). We therefore refer to this set of nurse observables as 'hiring characteristics': information on the nurse's ethnicity, age, qualifications, training and previous experience which a hiring manager would be able to observe when recruiting her into the FNP.

The FNP IS data also contains detailed measures of mothers' characteristics at programme intake.[9] Once mothers and nurses have been matched, they together complete a set of detailed intake paperwork. These forms cover the mother's health (including health conditions, height and weight); health habits (recent smoking, drinking and drug use); relationship (history of domestic violence); and demographics (including household composition, education, employment and income, and current use of health and social care services).

3.3 Data on visits and process quality

We also have two sources of 'process quality' data which provide information about how the nurse performs on the job. The first set of monitoring data is recorded by the nurse herself: after every home visit she conducts, she records the visit duration and what share of the planned content was covered; the proportion of time spent on each of the five programme content strands; and rates the mother's engagement in the visit and her understanding of and conflict with the material on three five-point scales (see Figure 2 and Figure 3). Importantly, the nurse is not asked to rate her own performance directly, but is instead prompted to record a mix of objective measures and assessments of how the visit went *for the mother*; this might have encouraged nurses to report these measures more honestly.

Even so, these monitoring data could suffer from reporting bias in unpredictable ways (for example, less skilled nurses might give overly positive reports from their visits, but they might also put greater weight on negative interactions within the visit). We therefore make use of a second source of 'process quality' data, taken from supervisors' evaluations of the

[9] We do not have access to information from mothers' referral forms, but most of this light-touch information on due date, language, etc. was re-collected at intake.

nurse. As discussed in section 2, supervisors occasionally accompany the nurse they manage on one or more of her home visits and assess the nurse against the Visit Implementation Scale.[10] This thorough evaluation takes in all aspects of preparing for and conducting the visit, including planning/scheduling; greeting and relating to the client; reviewing previous content and goals; assessing the client's current status; delivery of planned curriculum; and goal setting for the next visit.

While the supervision visits did have high stakes for nurses' careers, in Figure 6 we show that having a supervision did not affect nurses' practices, either at the time of the supervision or in the visits she conducted subsequently.[11] There is a clear trend in the proportion of time spent on maternal health and on child development content (reflecting the expected shift in programme content as a nurse's caseload ages); however, there is no discontinuous effect either in the month of the supervision or in the period following.

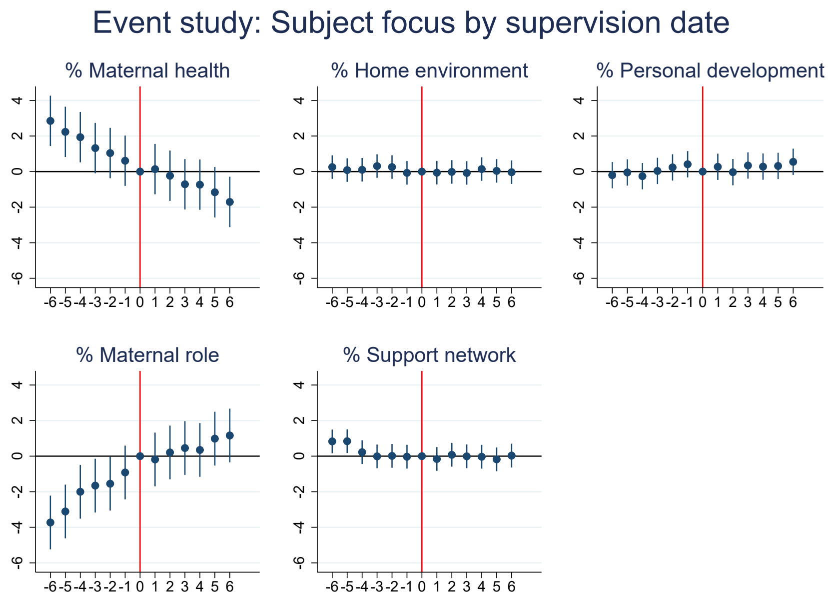
4 Empirical strategy

Our aim in the first part of the paper is to estimate the overall impact or effectiveness of each family nurse on children's outcomes at different ages. To explain the empirical strategy that we will use, in this section we start by sketching a simple dynamic model of skill formation following Todd and Wolpin (2003). The model helps formalize the identification assumptions underlying our empirical analysis and provides an interpretation of the parameters that we estimate. We then provide supportive evidence for our identification assumptions.

[10]These supervisory visits were intended to occur roughly every six months, but in practice we find considerable heterogeneity in their frequency and timing.

[11]The Figure shows the (lack of) impact that supervisions have on visit content, since these are the more objective nurse-reported process quality measures. However, we also find no impact on the other process quality measures recorded by nurses (results available on request).

Figure 6: Visit content by time to/since a supervision



Note: The figure shows the coefficients from event study regressions of visit content on the months to/since a nurse has a supervision visit. We exclude supervisions where the nurse has another supervisory visit within 6 months. Source: Authors' calculations using data provided by the Department for Health and Social Care.

4.1 Conceptual framework

As in the recent literature on child development, we model human capital as a multidimensional object, which evolves over time as a function of a series of inputs. Let θ_{ina} denote the human capital of child i at age a assigned to family nurse n in team s. θ_{ina} is a multi-dimensional vector, reflecting the different components of human capital, such as cognition, socio-emotional skills and health (Cunha and Heckman, 2008; Cunha et al., 2010). Following the framework of Todd and Wolpin (2003), we conceive human capital accumulation as a production process in which current and past inputs are combined with the individual's genetic endowments (determined at conception) to produce a particular level of human capital. We adapt the framework of Todd and Wolpin (2003) to make it relevant for children participating in the FNP and assume that inputs reflect choices made either by the mother or by the family nurse n assigned to the child. We denote the vector of mother-supplied inputs at a given age a as M_{ina}, family nurse-supplied inputs N_{ina}, and the vectors of their respective input histories up to age a as $M_{in}(a)$ and $N_{in}(a)$. Further, we let the child's genetic endowment be denoted as α_i, which we implicitly assume to be uni-dimensional.

Given these assumptions, the evolution of the child's human capital over time can be summarized by the following production function:

$$\theta_{ina} = g_a(M_{in}(a), N_{in}(a), \alpha_i) \tag{1}$$

We index the production function g with the a subscript to indicate that the impact of inputs and of the genetic endowment are allowed to depend on the age of the child.

Human capital θ_{ina} is rarely observed perfectly. Rather, analysts observe test scores and health indicators, which can be considered as imperfect proxies of underlying dimensions of human capital. To reflect this, we re-write the production function above so that the outcome is a particular measure of human capital T_{ina}, possibly measured with error:

$$T_{ina} = f_a(M_{in}(a), N_{in}(a), \alpha_i, \xi_{ina}) \tag{2}$$

As elaborated in Todd and Wolpin (2003), there are several problems when empirically implementing the model specified in equation (2). The first one is that the genetic endowment α_i is non-observable. The second one is that data sets on inputs are incomplete; despite rich data from the FNP IS, we do not observe the entire history of inputs supplied by the mother and by the family nurse.

To explain how we deal with these issues and for expositional clarity, let us assume that the production function can be linearly approximated and that only a set of mother's characteristics at programme's intake are observed. Specifically, we rewrite the production function above as follows:

$$T_{ina} = \beta_0 + X_{int_0}\beta_1 + \mu_n + \epsilon_{ina} \tag{3}$$

where T_{ina} is the test score of child i visited by nurse n at age a. X_{int_0} is a vector of characteristics of the mother at programme intake, and μ_n is a family nurse fixed effect proxying for all time-invariant characteristics of FN n; ϵ_{ina} is the error term. In this framework, the parameter μ_n or family nurse fixed effect captures the total effect of FN n on child i that is common to all mother-children pairs assigned to her. In principle, it is possible that the impact of the family nurse depends on the child's characteristics; we assume this possibility away at present.

To tackle identification issues, it is useful to decompose the error term ϵ_{ina} in equation 3 above as follows:

$$\epsilon_{ina} = \alpha_i + \xi_{ina} + e_{ina} \tag{4}$$

where ξ_{insa} is a measurement error and e_{insa} is an idiosyncratic shock.

Before delving into identification issues, it is important to clarify the interpretation of the parameter μ_n in equation (3) above. This 'family nurse effect' captures the total effect of the family nurse. In particular, it captures her direct effect on the child; her indirect effect on the child (through changes in mother-supplied inputs); and any indirect impact that the FNP site, supervisor or wider team has on the child's outcomes. For example, the family nurse could affect children's outcomes both by directly interacting with the child, but also by providing the mother with information on good parenting practices, helping her quit unhealthy behaviours (smoking, drinking) and/or supporting her mental health. Given that in the FNP, the family nurse works in direct contact with both the mother and the child, it is likely that she impacts maternal inputs that are relevant for the child. In what follows, we discuss the identification assumption of μ_n in the model above, where the fixed effect μ_n is interpreted as a measure of overall effectiveness of family nurse n.

Hence, μ_n is the total effect of family nurse n on child outcome T_{insa} if the following condition holds:

$$E\left(\epsilon_{ina}|\mu_n, X_{int_0}\right) = 0 \tag{5}$$

There are several cases in which this assumption would break down. First, the assumption would not hold if family nurse quality is correlated with the measurement error ξ_{ina} or with the random shock e_{ina}: we assume away these possibilities for the time being. The second case where identifying the family nurse fixed effect in equation (3) is problematic is if the family nurse's unobserved characteristics μ_n are systematically correlated with the child's unobserved characteristics α_i. Such correlation could occur, for example, because supervisors purposefully match family nurses with particular training and skills to pregnant mothers with particular needs or characteristics. If that is the case and these maternal characteristics are correlated with the child's unobserved endowment (as we would expect them to be), it is likely that μ_n would be correlated with α_i. This issue has a clear parallel in the education literature, where teachers are not randomly sorted into classrooms in most schools.

In the teacher quality literature, researchers have dealt with this issue either by using experiments in which teachers are randomly (voluntarily or involuntarily) assigned to students (Kane et al., 2008; Deming, 2014; Angrist et al., 2017) or by using observable variables – specifically one or several lagged measures of child's attainment – to proxy the child's unobserved endowment and unobservable input history. The latter approach is the so-called "value-added specification", whereby students' test scores are regressed on contemporaneous inputs (if available), one or several lagged achievement measures and a vector of teacher fixed effects. In this formulation of the model, the teacher fixed effect captures the impact of the teacher on her students' achievement gains since the previous period. A crucial assumption in the value-added approach is that lagged achievement is a sufficient statistic for unobserved input histories as well as the unobserved genetic endowment. Given the obvious difficulty to randomly assign teachers to students in reality, the value-added approach is by far the most common in the economics and education literatures, and it has been implemented in various contexts to estimate teacher effectiveness on academic gains (see Koedel et al. (2015) for a review) and more recently on adult outcomes (Chetty et al., 2014a,b).

In this paper we overcome this identification issue by exploiting a unique feature of the FNP assignment process of the family nurses to the clients. As discussed in section 2, the appproach supervisors follow to assign nurses to new clients is known, and we can use the rich data at our disposal to account for it. Thus, conditional on the variables that define the assignment process, we can assume that μ_n is independent of α_i – or, in other words, that family nurses are randomly assigned to mothers. In what follows, we describe the assignment rule in detail and provide supportive empirical evidence that, conditional on the variables that govern it, FNs observable characteristics are systematically uncorrelated with mothers' characteristics.

4.2 Assignment of nurses to clients

As we discuss in section 2, mothers joining the FNP provide an initial set of basic information via their referral form before being allocated a family nurse by the FNP site supervisor. This allocation process follows a known set of rules driven by the caseload of nurses in the site, with potential adjustments for language and geographic location. Detailed information about the mother is not collected until she and her assigned nurse complete the intake paperwork together.

When assigning a family nurse to a new client, the FNP site supervisor follows a set of rules aimed at ensuring a balanced caseload across family nurses within a team. First, the supervisor analyses each nurse's caseload relative to that of other nurses within the team to identify nurses with capacity to take on another client. Second, she analyses the complexity of each nurse's existing caseload. For example, she considers whether some nurses have more safeguarding work than others, more mothers who do not speak English, or more mothers with physical disabilities or additional learning needs.[12] Third, she considers the geographical location of new clients relative to that of existing clients, as it might be easier for FNs to reach clients who are geographically close. Fourth, she considers the mother's due date, since FNP guidelines recommend that no more than four babies are born each month on a nurse's caseload.

4.2.1 Testing the quasi-random assignment of family nurses to mothers

Our identification strategy makes use of the fact that, when allocating a family nurse to a client, the supervisor's information set only includes the limited set of characteristics reported in the client's referral form on the one hand, and the caseload of the nurses in her team on the other. Thus, conditional on her relative caseload at the time of assignment, the allocation of a family nurse to the mother should be as good as random.

While this assumption is not directly testable, we provide empirical evidence of its validity by testing whether a wide array of family nurse characteristics are correlated with characteristics of the new client, after we condition on the month of enrollment and on a small set of demographic variables typically collected on referral forms (including the mother's age in months, her gestational age, her ethnicity and whether she speaks English). We consider both the nurse's characteristics (demographics, training, experience and so on) and the information she reports from previous visits with her existing caseload (such as visit duration, content coverage, and assessments of mothers' engagement).

Specifically, we estimate the following equation:

$$Z_{int_0} = \alpha + \beta B_{it_0} + \delta E_{it_0} + \gamma X^{CC}_{nt_{-1}} + \phi_{t_0} + \epsilon_{int_0} \tag{6}$$

where Z_{int_0} is a characteristic of family nurse n allocated to client i as measured in the

[12] FNP staff have repeatedly stressed to us that inferring a mother's actual level of vulnerability is very difficult, even with the extensive information collected once she and her assigned nurse complete the intake paperwork. FNP staff report that cases which might seem relatively straightforward in the first place frequently turn out to be more complex than expected.

month of the mother's enrollment (t_0). B_{it_0} and E_{it_0} are two vectors of characteristics of the new client i, both also measured at the time of enrollment t_0. B_{it_0} captures the client's basic demographics (the characteristics most commonly collected on referral forms): her age, whether she speaks English, her ethnicity, her partnership status, and her gestational age. We assume that these characteristics are in the supervisor's information set at the time of assignment and so we do not test whether they are correlated with the assigned nurse's characteristics. E_{it_0} captures the mother characteristics which are subsequently collected at intake (and which we expect to be uncorrelated with nurse characteristics in the presence of quasi-random assignment). Finally, as outlined above, we control for two measures of the nurse's pre-existing relative case complexity ($X^{CC}_{int_0}$) as well as for year-month dummies (ϕ_{t_0}).

Given that the supervisor uses the information collected on the new client's i at the time of referral exclusively for the purpose of balancing the caseload across team members and does not even have access to the more detailed information in E_{int_0} at the time of assignment, we should find no systematic association between family nurse's characteristics Z_{int_0} and the new client's characteristics E_{int_0} once we have conditioned on the client's basic demographics B_{it_0}, the nurse's existing relative caseload complexity $X^{CC}_{int_0}$, and the year and month of enrollment.

We note that equation (6) does not contain any site or team fixed effects. We observe relatively few family nurses switching sites over time; even among those who do serve multiple sites, 86% see only a single mother at their secondary site(s). This means we cannot disentangle family nurse effectiveness from the direct impact, if any, of the wider team on a child's outcomes. Since we will not be able to condition on team fixed effects when estimating nurse effectiveness, we impose a more stringent requirement on our checks of random assignment than is implied by the policy environment: we require quasi-random assignment to hold across *all* clients who join the FNP in a given year and month, regardless of their local site. This more stringent criterion means that our tests of the plausibility of quasi-random assignment might fail if assignment within a site is random but there is spatial correlation in both client and nurse characteristics. In practice, as we will see from the results below, the extremely strict national FNP eligibility criteria mean that clients from around the country are similarly deprived and so there is relatively little spatial correlation within site.[13]

Results We estimate equation (6) 37 times, using each of the 37 nurse characteristics in Z_{int_0} as the outcome. For each of these nurse characteristics, we test the joint predictive power of the mother's characteristics (E_{int_0}) using an F-test. These results are reported in Column 3 of Table 3 for the sample of children with birth outcomes.[14] For convenience, we group the 37 nurse characteristics into five categories and list the number of regressions in each category in Column 2.

Overall, we find that the mother's characteristics are significant predictors of three of the characteristics of her assigned nurse (in this case, the nurse's history of visit length; percent of content covered; and share of the visit spent on maternal physical health). The lack of a significant relationship between the mother's characteristics and the nurse's demographics, qualifications, training or experience strongly suggests that nurses are indeed quasi-randomly assigned on these dimensions.

Finding that three of the nurse process quality characteristics are significantly correlated with mother traits suggests that supervisors may take nurse process quality into account when assigning nurses to new mothers. However, this correlation may be spurious: all three of the nurse characteristics in question are strongly related to the average tenure of a nurse's caseload, which we do not control for. Further, while mother characteristics are jointly significant predictors of these three nurse traits, they are inconsistent ones. For example, nurses whose previous visits have lasted longer are more likely to be assigned to mothers with a history of social service use, but they are less likely to be assigned to mothers who

[13] This is borne out by the intra-class correlation in mother characteristics, which ranges between 0.001 and 0.066, with the higher ICCs found for characteristics related to housing tenure. In addition, site fixed effects are only poor predictors of mother characteristics: in a series of regressions of each mother characteristic on a full set of site fixed effects, the adjusted-R^2 ranged between 0.001 and 0.065, suggesting that spatial correlation within sites explains only a small share of the variation in mother characteristics in our sample.

[14] Results showing analogous tables for the analytical samples used for other outcomes are found in the Appendix.

have not achieved educational benchmarks.[15]

Table 3: Summary of tests for quasi-random assignment: Analytical sample for birth outcomes

Outcome group	# of outcomes	# of outcomes where mother Xs signficant		
		No MHT	*Holm-Bonferroni-Sidak*	*Modified Romano-Wolf*
Nurse demographics	5	0	0	0
Nurse qualifications	12	0	0	0
Nurse experience	5	0	0	0
Nurse training	5	0	0	0
Process quality	10	3	2	0
Overall number	37	3	2	0
Overall %		8.1%	5.4%	0.0%

Multiple hypothesis testing So far, we have implicitly assumed independence between the different mother characteristics and between between the individual nurse traits. This is clearly unrealistic: in fact, there is a high degree of correlation both within client characteristics, and within nurse traits. The simple approach we have applied so far is therefore ill-suited to the situation at hand, since we are likely to encounter multiple hypothesis testing issues along two dimensions: testing multiple correlated outcomes (nurse characteristics), and testing multiple correlated predictors (mother traits).

We therefore adapt the two principle MHT algorithms - the Holm-Bonferroni-Sidak adjustment, and the Romano-Wolf stepdown procedure - to the case where we have multiple predictors of interest as well as multiple outcomes. We report the number of nurse characteristics which are significantly correlated with mothers' characteristics under each of these

[15] Results available on request.

algorithms in Columns 4 and 5 of Table 3.

After correcting for multiple hypothesis testing, we find that 2 of our 37 nurse characteristics are significantly predicted (at the 5% level) by mother traits under the Holm-Bonferroni-Sidak adjustment. This is just over 5% of the outcomes. When we apply the more conservative Romano-Wolf algorithm, we no longer detect any significant relationship between the characteristics of the mother and the traits of the nurse who is assigned to her.

While this is not a formal test of the random assignment of family nurses to mothers, we interpret these results as strongly suggestive evidence that nurses are indeed quasi-randomly assigned to clients, conditional on their caseload complexity at the time of enrollment and the basic maternal demographics collected on the referral form.

5 Heterogeneity in family nurse effectiveness

Based on the conditionally random assignment of nurses to new clients outlined in section 4, we are now ready to estimate the degree of heterogeneity in family nurses' impacts on children's outcomes. As discussed in section 3, we consider five child outcomes: a factor of health at birth; cognitive scores on the Ages and Stages Questionnaire-3, collected at ages 1 and 2; and socio-emotional scores on the ASQ-Socio Emotional, again collected at ages 1 and 2. To ensure that the scale of nurse effectiveness is comparable across outcomes, all have been standardised within sample to have a mean of zero and a standard deviation of one.

To do this, we estimate the following equation:

$$Y_{int_a} = \beta B_{it_0} + \gamma X^{CC}_{nt_{-1}} + \phi_{t_0} + \mu_n + \xi_{int_a} \tag{7}$$

where Y_{int_a} is an age a outcome of child i whose mother enrolled at time t and who is visited by nurse n; B_{it_0} and $X^{CC}_{nt_{-1}}$ are the vectors of basic mother characteristics and nurse caseload complexity which we controlled for in equation (6) when assessing the random assignment; ϕ_{t_0} is a set of dummies for the mother's enrollment year and month; and ξ_{int_a} is an idiosyncratic error term. Our coefficient of interest is the nurse fixed effect μ_n. Note that we omit a constant so that we are able to estimate the nurse fixed effects for all family nurses in the sample.

Following Guarino et al. (2015) and much of the teacher value-added literature, we estimate μ_n using an empirical Bayes procedure, also known as a "shrinkage estimator". Nurses have different numbers of clients included in the estimation (ranging from 10 clients at the 10th percentile to 50 at the 90th); this means that some of the nurse fixed effects are estimated more precisely than others. The empirical Bayes procedure accounts for these differences by assigning greater confidence to the nurse fixed effects estimated with greater numbers of clients.[16]

5.1 Results

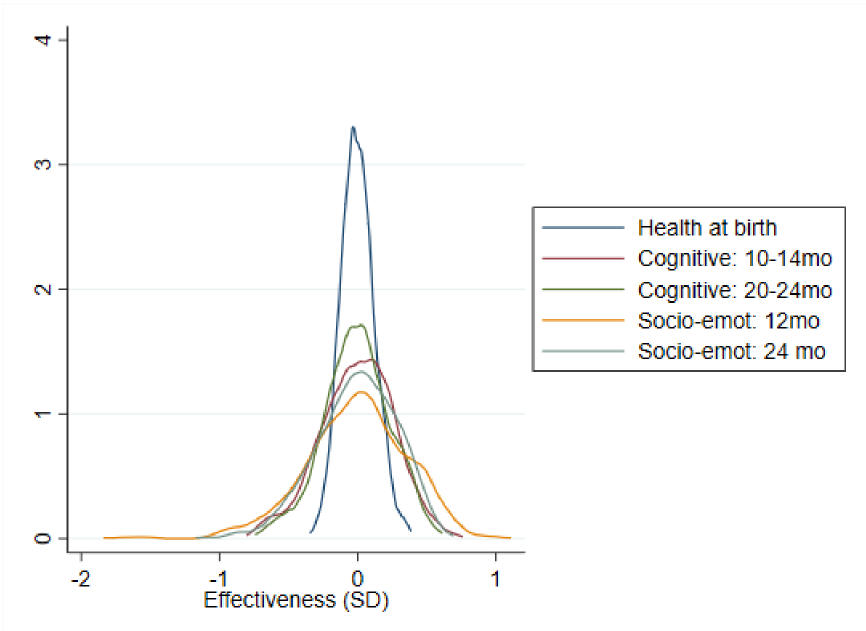
Table 4 reports various moments of the estimated distributions of the family nurse fixed effects for each of our five outcomes of interest. Figure 7 plots the distributions. Across outcomes, we find meaningful heterogeneity in family nurse effectiveness. The distribution is narrowest for children's birth outcomes, when mothers have had less exposure to their family nurse. We find that a one-standard deviation (SD) increase in family nurse effectiveness leads to a 0.121 SD improvement in our birth outcomes factor. For cognitive outcomes at both age 1 and 2, a one-SD increase in family nurse effectiveness improves outcomes by 0.24 to 0.27SD. We find the widest distribution for socio-emotional outcomes at age 1; here, a one-SD increase in family nurse effectiveness improves scores on the ASQ-SE by 0.37 SD. As Figure 7 shows, this wider distribution is driven by a much longer left tail of the distribution, indicating that less effective nurses have a larger impact on socio-emotional development.

Importantly, these outcomes are measured *relative* to the average nurse. We cannot conclude from this whether less effective nurses are actively *harmful* (i.e., children would

[16]Results that do not apply this shrinkage procedure are found in the Appendix.

have had better development without the FNP), or whether they simply deliver (much) smaller benefits for child development than their more effective peers.

Figure 7: Distribution of family nurse effectiveness



Note: The figure shows the distribution of family nurse value-added in promoting each of five child outcomes. Results are based on a sample of nurses who had joined the FNP by December 2013 and who had at least five clients with valid outcome measures. Sample sizes are reported in Table 4. Results have been adjusted using an empirical Bayes estimator ("shrinkage estimator") to account for differences in number of clients per nurse. Source: Authors' calculations using data provided by the Department for Health and Social Care.

5.1.1 Randomisation inference

In order to assess the economic significance of these results, it is important to have a sense of how wide the distribution of family nurse effectiveness is. That is, how likely would we be to observe a similarly wide - or wider - distribution of nurse effectiveness, even if there were no link between nurses and their clients' outcomes? While standard methods of inference based on test statistics do not extend easily to this case, we adopt a randomisation inference approach to assess whether our estimates suggest an 'unexpected' degree of heterogeneity in family nurse effectiveness.

Our randomisation inference test proceeds under the null hypothesis that there is no

Table 4: Distribution of family nurse effectiveness for different child outcomes

	Child's outcome				
	Birth outcomes	ASQ-3: age 1	ASQ-3: age 2	ASQ-SE: age 1	ASQ-SE: age 2
Standard deviation	0.121	0.266	0.239	0.371	0.292
10th percentile	-0.145	-0.358	-0.310	-0.459	-0.392
25th percentile	-0.079	-0.181	-0.156	-0.229	-0.192
50th percentile	-0.007	0.008	-0.008	0.013	0.009
75th percentile	0.075	0.182	0.142	0.231	0.208
90th percentile	0.166	0.324	0.218	0.449	0.358
p90-p10	0.311	0.682	0.628	0.908	0.750
p75-p25	0.154	0.363	0.298	0.460	0.400
(p90-p10)/SD	2.570	2.564	2.628	2.447	2.568
(p75-p25)/SD	1.273	1.365	1.247	1.240	1.370
Number of nurses	647	616	585	611	574
Number of mothers	18,258	13,820	10,383	13,060	9,700

Note: The numbers reported in each column are the Empirical Bayes estimates of the family nurse fixed effects.

systematic relationship between nurses' effectiveness and their clients' outcomes (and so any heterogeneity in the estimated nurse fixed effects are pure noise). We test this by constructing the distribution of nurse effectiveness when we randomly reshuffle children and nurses. While our main tests of quasi-random assignment of nurses in section 4 show that there are few if any systematic relationships between nurses and mothers at the *beginning* of the FNP, by randomly re-assigning nurses to children we also remove any impact that the nurse has had during the course of the programme. We then estimate a set of nurse fixed effects based on this randomised dataset, and compute the standard deviation of this distribution. We repeat the exercise 500 times for each of our five outcomes.

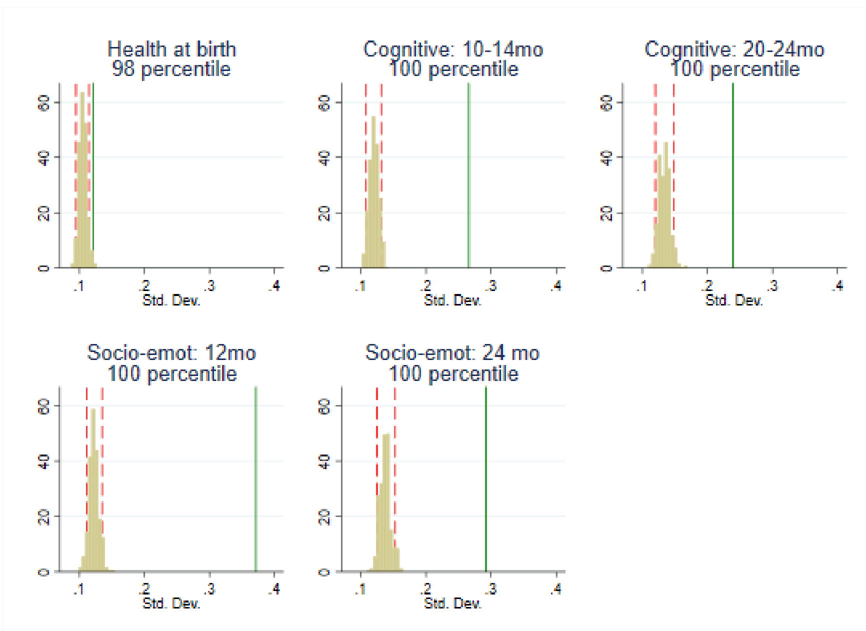
In Figure 8, we then plot (in beige) the distribution of the standard deviation of the nurse fixed effects arising from these 500 permutations. The red vertical lines show the 5th and 95th percentiles of the distribution - where the standard deviation of the nurse effectiveness distribution falls within these bounds, we cannot rule out estimating a similarly wide distribution even absent any actual differences in nurse effectiveness.

We show the true standard deviations with green vertical lines. In each case, this true degree of heterogeneity in family nurse effectiveness (measured by the true standard deviation of the distribution of nurse fixed effects) lies outside these bounds, indicating that the true degree of heterogeneity is larger than we would expect to observe by chance. However, we note that the true standard deviation of the distribution of nurse effectiveness for health at birth does overlap the tail of the randomisation inference distribution. This indicates that, although there is a statistically significant degree of heterogeneity, its magnitude is still small relative to the heterogeneity present due to pure noise.

We can also use the randomisation inference test to give us a sense of the magnitude of the heterogeneity we estimate. In Table 5, we report the median standard deviation from our permutation tests, the true standard deviation, and the ratio between them. This ratio captures how much heterogeneity in family nurse effectiveness we detect, compared to what would have been expected even if the true impact of nurses on children were zero. This exercise essentially lets us assess the degree of heterogeneity against a more realistic null hypothesis than assuming that the standard deviation of nurse effectiveness is zero.

We find that there is around 20% more heterogeneity in family nurses' effectiveness at improving birth outcomes than we would expect by chance. For cognitive outcomes, we find around twice as much heterogeneity as would be expected by chance. The greatest degree of heterogeneity comes in age 1 socio-emotional outcomes, where the nurse effectiveness distribution is three times as wide as we would otherwise expect. This is driven by the long left tail in nurse effectiveness for this outcome shown in Figure 7.

Figure 8: Randomisation inference test of the degree of heterogeneity in family nurse effectiveness



Note: In beige, we plot the distribution of the standard deviation of nurse fixed effects. This is constructed by randomly reassigning nurses to families and estimating the set of nurse fixed effects, then calculating the standard deviation of that distribution. We build up the distribution shown in beige through 500 permutations of this procedure per outcome. Red dashed vertical lines indicate the 5th and 95th percentiles of this randomisation inference distribution. Green vertical lines show the true standard deviation of the distribution of family nurse effectiveness (as in Table 4). Source: Authors' calculations using data provided by the Department for Health and Social Care.

Table 5: Summary of permutation-based and actual heterogeneity in nurse effectiveness

Outcome	Median permutation SD	True SD	Ratio (True/Permutation)
Birth outcomes	0.105	0.121	1.2
Cognitive, age 1	0.120	0.266	2.2
Cognitive, age 2	0.134	0.239	1.8
Socio-emot, age 1	0.122	0.371	3.0
Socio-emot, age 2	0.137	0.292	2.1

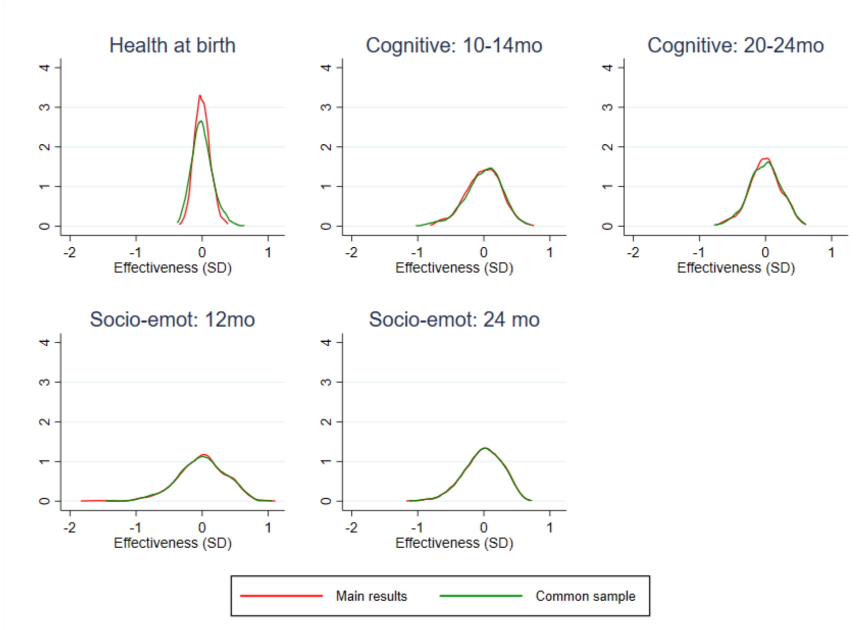
5.1.2 Robustness and sensitivity checks

In order to assess the robustness of our estimates of nurse effectiveness, we carry out a range of robustness and sensitivity checks.

Common sample In Figure 9, we show the impact of imposing a common sample across our five outcomes of interest. This common sample uses only children with valid results for all five outcomes to estimate nurse effectiveness (and so drops 73 nurses relative to our main estimates for birth outcomes). Imposing this common sample leads to a slightly wider distribution of nurse effectiveness for birth outcomes, which is consistent with a smaller number of available clients per nurse leading to less precision in our estimated fixed effects. The results for all other outcomes look similar to our main estimates.

Additional control variables We have argued that the assignment of mothers to family nurses was random conditional on the nurse's caseload, the time of enrollment and basic mother demographics. While we have provided strong suggestive evidence for this identification strategy, it remains fundamentally untestable. In this robustness test, we therefore control for a range of additional information about the mother, the supervisor and the local area when estimating a nurse's effectiveness. To the extent that our random assignment assumption is violated in ways that correlate with any of these variables, we would expect

Figure 9: Impact of imposing a common sample on the distribution of family nurse effectiveness



Note: See Figure 7. The 'common sample' results impose a common sample across all five outcomes, comprising 569 nurses. Source: Authors' calculations using data provided by the Department for Health and Social Care.

their inclusion to shift our estimates of nurse effectiveness.

We first control for the extended set of mother's characteristics at intake, E_{it_0}. We showed in section 4 that these characteristics were largely uncorrelated with the characteristics of the assigned family nurse; however, to the extent that the quasi-random assignment breaks down, including them here will correct for any omitted variable bias and so change our estimates of nurse value added. As shown in Figure 10, the results after inclusion of these mother controls (show in green) are virtually indistinguishable from our main results (in red).

Next, we control for characteristics of the supervisor. Specifically, we control for the same set of characteristics of the nurse that we used in assessing the random assignment (Z_{int_0}), but measured for the nurse's supervisor. These proxy for any unobserved differences between teams. Again, we find that the inclusion of these additional controls makes essentially no difference to our estimates of nurse effectiveness.

As a final set of controls, we include time-varying characteristics of the local authority.[17] We construct an extensive dataset of local authority controls, including local economic conditions (male and female median weekly earnings and the local claiming rate for unemployment benefits); local vital statistics (the fertility rate, teenage conception rate and proportion of children with low birth weight); local demographics (population density and the proportion of children who are non-native English speakers); social care use (the rate per 1,000 of children aged less than 1 and children aged 1 to 4 being looked after by the local authority); and local service provision (the take-up of universal free childcare places and the number of GPs per 1,000). These time-varying local characteristics reflect some of the geographic differences between sites, and hence the potential differences between children's trajectories around the country.

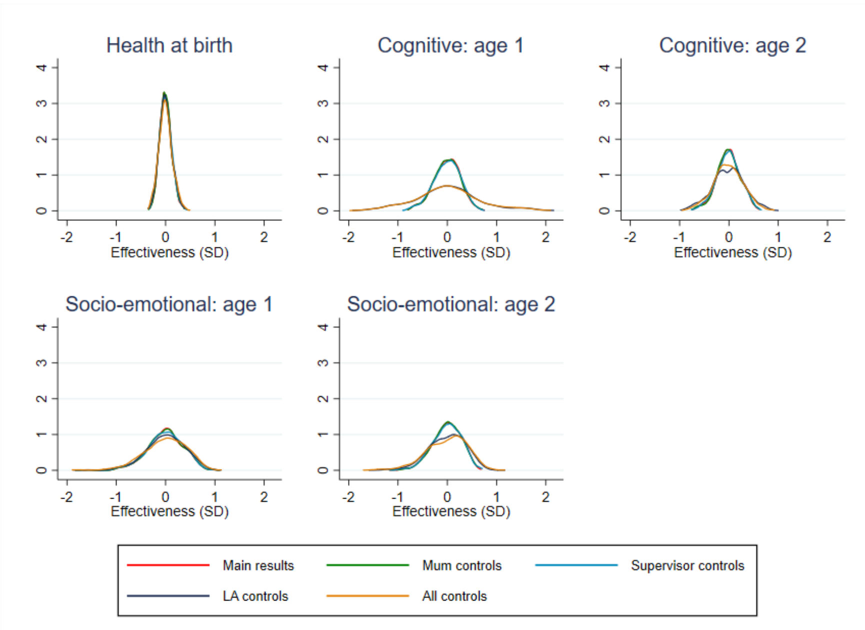
For most of our outcomes, we find that the inclusion of local area characteristics also makes relatively little difference to our main estimates (blue). However, these controls do substantially change the distribution of nurse effectiveness in promoting cognitive outcomes, especially at age 1. This is driven by the inclusion of controls for population and population density, and likely reflects a well-documented 'London effect' whereby children from deprived backgrounds tend to enjoy much better educational trajectories if they live in major urban areas.

Finally, we estimate a specification that includes all three groups of additional controls (in yellow). We find that this mirrors the distribution of the specification with local area controls almost perfectly.

Sensitivity to different samples In our main analysis, we restrict the analytical sample to nurses and mothers who joined the FNP before 2014. This helps to ensure that mothers have had time to complete the programme by the time our data ends, in February 2017. We

[17] These are measured in a yearly panel at the level of the lower-tier local authority, of which there are 326 in England.

Figure 10: Robustness of family nurse effectiveness estimates to the inclusion of additional control variables

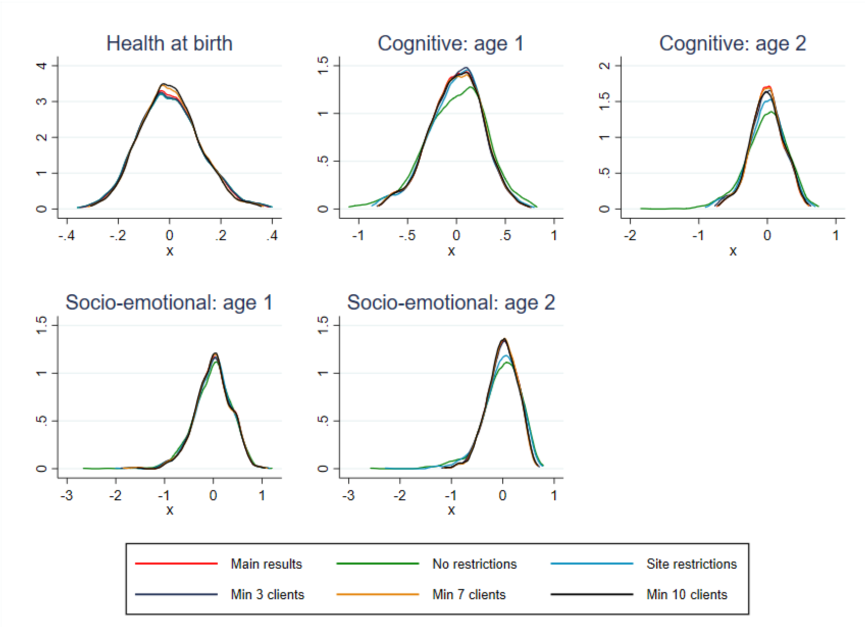


Note: See Figure 7. 'Random assignment' controls include the 37 mother characteristics assessed in our random assignment checks. 'Supervisor' controls include the set of nurse characteristics used in our random assignment checks, but measured for the supervisor rather than the family nurse. 'LA controls' include time-varying controls at the local authority level, measured at the time of intake. These include local economic conditions (male and female median weekly earnings and the local claiming rate for unemployment benefits); local vital statistics (the fertility rate, teenage conception rate and proportion of children with low birth weight); local demographics (population density and the proportion of children who are non-native English speakers); social care use (the rate per 1,000 of children aged less than 1 and children aged 1 to 4 being looked after by the local authority); local service provision (the take-up of free childcare places and the number of GPs per 1,000); and spending on services for children and young people. Source: Authors' calculations using data provided by the Department for Health and Social Care.

further require that nurses have at least five valid clients to be included in the sample, to improve the precision of the estimated fixed effects.

In Figure 11, we analyse how sensitive our estimates of nurse effectiveness are to these decisions. We largely find few differences between the various specifications. Where there are differences, our main estimates yield a slightly narrower distribution of nurse effectiveness than in specifications that do not impose a minimum number of clients per nurse. This is consistent with the greater precision that motivated us to introduce a minimum number. However, we find no evidence that imposing higher thresholds changes the estimates of nurse effectiveness. The threshold of 5 or more clients per nurse that we use in our main results therefore strikes a balance between improving precision without unnecessarily discarding nurses from the analysis.

Figure 11: Sensitivity of family nurse effectiveness estimates to different analytical samples



Note: See Figure 7. To reduce noise in the estimates of effectiveness, our main results exclude any nurse with fewer than 5 valid clients and require all valid clients to have joined the FNP by December 2013. The results with no restrictions ('none') do not impose either restriction. The other sensitivity checks impose the pre-2014 joining restriction, but vary the minimum number of clients required for a nurse to be included in the estimation. The figure is based on a common sample across all six specifications. Source: Authors' calculations using data provided by the Department for Health and Social Care.

5.2 Correlation between different domains of effectiveness

We estimate nurse effectiveness separately for each of our five child outcomes of interest, allowing nurse effectiveness to vary freely between outcomes. In this section, we explore the extent to which nurse effectiveness is correlated across outcomes. That is, are family nurses better thought of as specialists (where an individual nurse is outstanding in promoting one dimension of child development, but average or below-average in fostering other domains)? Or are they more like generalists, where higher effectiveness applies to all domains of child development?

We therefore compute the correlation between the family nurse fixed effects estimated for each of our five outcomes of interest. These correlations are reported in Table 6. Overall, we find that nurse effectiveness is correlated across most domains of child development. As expected, the correlations are particularly strong within developmental domains - for example, there is a 69% correlation between a family nurse's effectiveness in promoting cognitive development at age 1 and at age 2. However, there is significant cross-domain correlation as well: the correlation in nurse effectiveness across cognitive and socio-emotional outcomes is around 30-40%. We find much weaker correlations, if any, between family nurse effectiveness in promoting better health at birth on the one hand, and better cognitive and socio-emotional development on the other. Even here, though, all correlations are positive, suggesting that family nurses who are good at improving one outcome are also systematically good at improving another outcome.

Table 6: Correlation between nurse effectiveness in promoting different child outcomes

	Birth	Cognitive		Socio-emotional	
	(1)	(2)	(3)	(4)	(5)
(1) Birth outcomes		**0.141***	**0.111***	0.008	0.066
(2) Cognitive, age 1	**0.141***		**0.691***	**0.390***	**0.389**
(3) Cognitive, age 2	**0.111***	**0.691***		**0.316***	**0.432***
(4) Socio-emot, age 1	0.008	**0.390***	**0.316**		**0.611***
(5) Socio-emot, age 2	0.066	**0.389***	**0.432***	**0.611***	

6 What makes a good family nurse?

So far, we have shown that there is a substantial degree of heterogeneity in family nurses' effectiveness at improving child development, particularly in respect of cognitive and socio-emotional development. This means that - even within a standardised programme delivered by a highly trained, highly paid workforce - there are substantial developmental returns to being assigned a more effective family nurse. Recruiting, retaining and training effective family nurses therefore has implications for the efficacy of the intervention as a whole.

In this section, we therefore explore the predictors of higher family nurse effectiveness. We exploit the rich data on nurses held in the FNP Information System to consider three groups of predictors. First, we assess the predictive power of the nurse's demographics, level and field of education, training and previous experience. We add to this two 'match quality' predictors: the proportion of a nurse's caseload who match her ethnicity (ethnic minority or not), and the proportion of a nurse's caseload who match her experience (aged 17 or younger for nurses with experience working with teen parents, or 18 and up for nurses without). We refer to these as 'hiring characteristics' because they cover objective, observable information that would commonly be available to a hiring manager on a CV. Consistent with much of the existing literature on teacher and workforce value-added, we find that these observable characteristics explain very little of the variation in nurse effectiveness.

The value-added literature has identified 'process quality' as a more robust metric to assess worker effectiveness. In our context, this is analogous to the assessments carried out by supervisors accompanying family nurses on their home visits. As outlined in section 3, the supervisors record detailed assessments of the family nurse across all areas of competence; in total, they record their judgement on 44 different criteria.

The final set of characteristics we consider is the visit metadata that nurses record after each visit. This covers the length of the visit; the proportion of planned content covered; the proportion of time spent on content from each of the five domains of the FNP programme; and assessments of how engaged the mother was, how well she understood the content, and the extent to which she showed 'conflict' with the content. We view these self-assessments as containing information not only about the mother, but also about her family nurse: since building relationships with mothers and presenting the material in a way that they feel engaged with is a core part of the family nurse role, low ratings in these areas indicate that the nurse has not managed to overcome the challenges of working with that client.

6.1 Correlates of family nurse effectiveness

In order to understand how well these three groups of characteristics predict family nurse effectiveness, we first assess the pairwise correlation between our estimated effectiveness and family nurse characteristics. We focus on nurses' effectiveness at improving children's cognitive development at age 2, since this is a terminal outcome of the programme.[18]

Figure 12 shows that there is very little association between a family nurse's observable characteristics and her effectiveness. Family nurses with experience working as a midwife or as a school nurse tend to be somewhat more effective, while those who have done some post-graduate study tend to be somewhat less effective. But in all cases these correlations are relatively small.

Figure 13 continues this exercise by looking at the correlations between family nurse effectiveness, process quality measures based on supervisor assessments and visit metadata. To reduce dimensionality, we aggregate the 44 different supervisor scores by taking the total score for each of 10 sub-categories of the assessment. Despite the predictive power of independent assessments of process quality in other contexts, we find that supervisors' assessments hold little information about nurse effectiveness; the ten assessment domains are not jointly significant predictors of family nurse effectiveness, and none of them individually correlates significantly with effectiveness. Indeed, most of the correlations we estimate are negative.

By contrast, the visit metadata recorded by family nurses after each visit is more pre-

[18]We find similar results when we instead focus on family nurse effectiveness in promoting socio-emotional development at age 2 (results available on request).

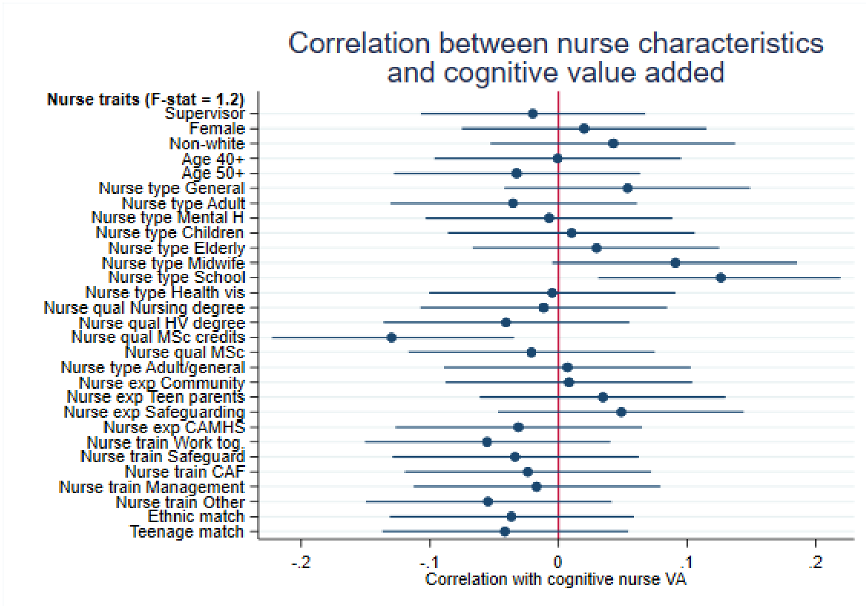
dictive of the family nurse's effectiveness. Nurses who report covering more content, having longer visits, and seeing better engagement and understanding from the mothers are on average more effective at boosting the child's development. This correlation likely reflects a number of channels. There may be a causal effect if, for example, longer visits, more content delivery and a better relationship with the mother mean more scope for the nurse to benefit the child. The correlation may also be picking up differences in the quality of the nurse-family match or in the child's unobserved ability. While our quasi-random assignment suggests that, on average, nurse characteristics are unrelated to family traits, some nurses will randomly end up with a caseload with higher-than-average ability, which will affect the child's outcomes and measured nurse effectiveness. This drives the non-zero heterogeneity in nurse effectiveness in our randomisation inference tests (Figure 8).

Regardless of the extent to which these visit metadata questions capture a causal impact of the nurse or simply reflect the match quality or family circumstances, our results suggest that straightforward, relatively low-cost 'monitoring' information can be at least as informative as formal assessments by a supervisor in identifying families and practitioners who may need more support to receive the full benefit of the programme.

So far we have focused on simple correlations between individual predictors and family nurse effectiveness. However, since these predictors are also correlated with one another, this gives an incomplete picture of what the most predictive characteristics might be. We therefore run nurse-level regressions of family nurse effectiveness for each of our outcomes on all of the characteristics in Figures 12 and 13, as well as the local area characteristics used in Figure 10. We then implement a Shorrocks-Shapley decomposition to partition attribute the proportion of the variance explained by each category of predictors. Because each group contains a different number of predictors, we focus on the adjusted R^2 to strip out the mechanical impact on the R^2 of having more independent variables.

The results in Figure 15 show that, even with this exceptionally rich set of potential predictors, we can explain only a small share of the variation in estimated family nurse

Figure 12: Pairwise correlations between family nurse effectiveness in promoting age 2 cognitive development and hiring characteristics



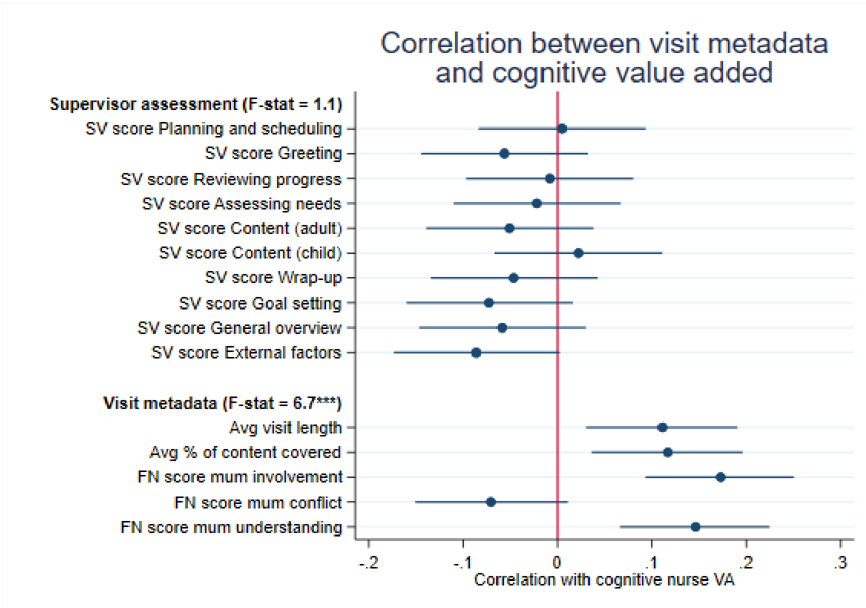
Note: See Figure 7. The figure shows the point estimate and 95% confidence intervals for the pairwise correlation between family nurse effectiveness and each demographic characteristic. The F-statistic comes from a regression of family nurse effectiveness on all characteristics in the figure. Due to missingness in nurse characteristics, these results are based on 419 nurses. Source: Authors' calculations using data provided by the Department for Health and Social Care.

effectiveness (below 15% when using the R^2, or less than 5% when using the adjusted R^2, our preferred measure). In absolute terms, the set of observable nurse 'hiring' characteristics (in blue) explains the greatest share of the variance in estimated family nurse effectiveness (left panel). However, this is due to the larger number of individual variables in this group; when looking at the adjusted R^2 (right panel), the visit metadata clearly dominates the other two groups of characteristics. Strikingly, information collected from supervisor assessments has very low predictive power for any of the outcomes we consider.

7 Conclusion

There is a large and rich evidence base on the impacts of early interventions on children's outcomes. While there is compelling evidence that early interventions *can* have substantial and

Figure 13: Pairwise correlations between family nurse effectiveness in promoting age 2 cognitive development and process quality and metadata measures

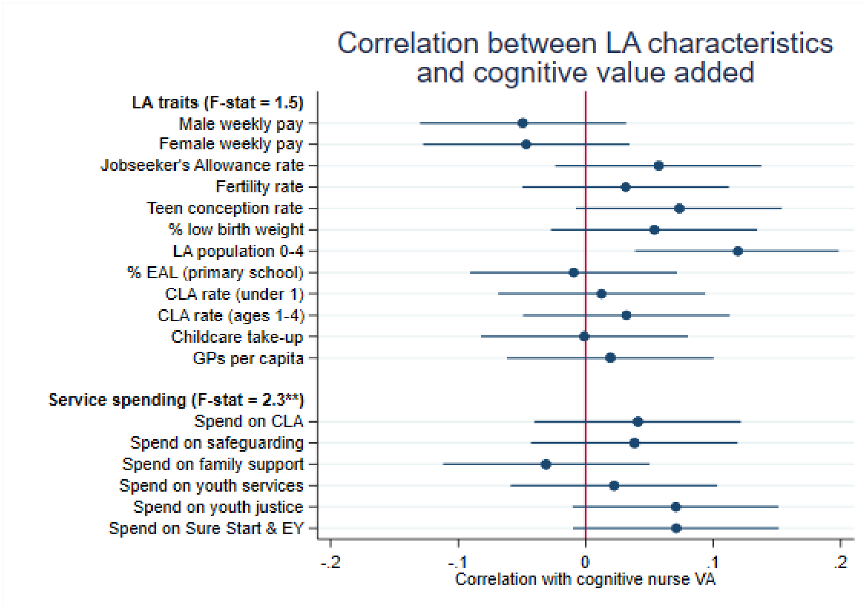


Note: See Figure 12. Results are based on 485 nurses. Source: Authors' calculations using data provided by the Department for Health and Social Care.

long-lasting benefits, there is also a large volume of research finding that some interventions fail to live up to their potential, particularly when they are delivered at scale. Workforce quality and effectiveness has been identified as a crucial component of successful early interventions, but so far very little is known about the degree of heterogeneity in workforce effectiveness or about the predictors of more effective workers.

In this paper, we have begun to fill this gap by showing that there is a substantial degree of heterogeneity in workforce effectiveness. We conclude that a one-standard deviation increase in family nurse effectiveness improves children's cognitive development at age 2 by 0.24 standard deviations, and improves socio-emotional development at age 2 by 0.29 SD. This is around twice the degree of heterogeneity that would be expected even if family nurses had no real effect on children's outcomes. Further, there is a reasonably high degree of correlation between nurse effectiveness at improving cognitive and socio-emotional development. This indicates that some nurses have an absolute advantage in improving both domains of child

Figure 14: Pairwise correlations between family nurse effectiveness in promoting age 2 cognitive development and local area characteristics

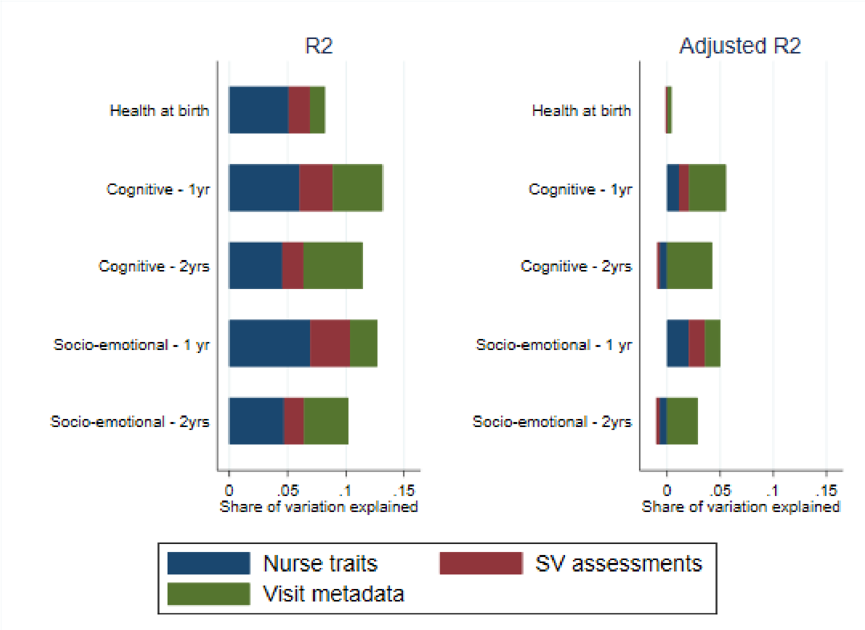


Note: See Figure 12. Results are based on 585 nurses. Source: Authors' calculations using data provided by the Department for Health and Social Care.

development. By contrast, we find a much more modest degree of heterogeneity in nurse effectiveness at improving birth outcomes, and little to no correlation with effectiveness at improving other outcomes.

Despite access to an unusually rich set of potential predictors of family nurse effectiveness, we find that we are only able to explain a modest share of this heterogeneity - less than 15% of the total variation in effectiveness. In common with much of the teacher value-added literature, we find that the characteristics observed during the hiring process - training, experience, demographics and so on - are barely predictive of family nurse effectiveness. More surprisingly, we also find very little relationship between a nurse's effectiveness and the assessments made by her supervisor. Instead, we find that the visit implementation data collected by the family nurse herself - including both objective measures like the duration and more subjective assessments of the mother's engagement - are significantly correlated with nurse effectiveness. This suggests that collecting similar metadata could be an important tool

Figure 15: Shorrocks-Shapley decomposition of the predictive power of different groups of family nurse characteristics for estimated family nurse effectiveness



Note: We implement a Shorrocks-Shapley decomposition for the R^2, left, and adjusted R^2, right. 'Nurse traits' contains the nurse's observable 'hiring' characteristics.'SV assessments' refers to the subdomain-specific supervisor assessment scores. 'Visit metadata' contains nurse-level averages of visit metadata (e.g. duration or mother's understanding). Source: Authors' calculations using data provided by the Department for Health and Social Care.

for monitoring the implementation of interventions and identifying families and practitioners with need of greater support, in close to real time.

These results are particularly notable because they come in the context of a highly structured programme that has been successful overall at improving children's development, and in the context of a workforce which is highly educated, highly trained and highly paid. These findings cast some doubt on the ability of formalised curriculum and a more highly trained workforce - both common proposals to promote more uniform delivery interventions - to fully overcome the role of differences between providers in shaping the effectiveness of early years programmes. Overall, our results show that the quality of the workforce matters, and that we are just starting to understand its determinants.

References

Almond, D. and J. Currie (2011). Human capital development before age five. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics, Volume 4b*. Amsterdam: Elsevier Science.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017). Leveraging Lotteries for School Value-Added: Testing and Estimation. *The Quarterly Journal of Economics 132*(2), 871–919.

Araujo, M. C., P. Carneiro, Y. Cruz-Aguayo, and N. Schady (2016). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics 131*(3), 1415–1453.

Avellar, S. and L. Supplee (2013, 11). Effectiveness of home visiting in improving child health and reducing child maltreatment. *Pediatrics 132 Suppl 2*, S90–9.

Blanden, J., K. Hansen, and S. McNally (2017, February). Quality in Early Years Settings and Children's School Achievement. CEP Discussion Papers dp1468, Centre for Economic Performance, LSE.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a, September). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b, September). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–79.

Cunha, F. and J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F., J. Heckman, L. Lochner, and D. Masterov (2006). Interpreting the evidence on life cycle skill formation. In E. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education, Volume 1*. Amsterdam: Elsevier Science.

Cunha, F., J. Heckman, and S. Schennach (2010). Estimating the technology of cognitive and non-cognitive skill formation. *Econometrica 78*(3), 883–931.

Deming, D. J. (2014, May). Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review 104*(5), 406–411.

Doyle, O., C. Harmon, J. Heckman, C. Logue, and S. H. Moon (2013). Measuring investment in human capital formation: An experimental analysis of early life outcomes. Working Paper Series, UCD Centre for Economic Research, No. 13/10.

Guarino, C. M., M. Maxfield, M. D. Reckase, P. N. Thompson, and J. M. Wooldridge (2015). An evaluation of empirical bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics 40*(2), 190–222.

Hanushek, E. A. and S. G. Rivkin (2006). Teacher quality. In *Handbook of the Economics of Education*, pp. 1051–1078.

Hanushek, E. A. and S. G. Rivkin (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics 4*(1), 131–157.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008, December). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review 27*(6), 615–631.

Koedel, C., K. Mihaly, and J. E. Rockoff (2015). Value-added modeling: A review. *Economics of Education Review 47*(C), 180–195.

Olds, D. (2015). Building evidence to improve maternal and child health. *The Lancet 387*(10014), 105–107.

Olds, D. L., C. R. Henderson Jr, H. J. Kitzman, J. J. Eckenrode, R. E. Cole, and R. C. Tatelbaum (1999). Prenatal and infancy home visitation by nurses: Recent findings. *The future of Children*, 44–65.

Peacock, S., S. Konrad, E. Watson, D. Nickel, and N. Muhajarine (2013, 01). Effectiveness of home visiting programs on child outcomes: A systematic review. *BMC public health 13*, 17.

Robling, M., M.-J. Bekkers, K. Bell, C. C. Butler, R. Cannings-John, S. Channon, B. C. Martin, J. W. Gregory, K. Hood, A. Kemp, et al. (2015). Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (building blocks): a pragmatic randomised controlled trial. *The Lancet 387*(10014), 146–155.

Salinas, P. and I. Armijo (2017). Predictive value of ages & stages questionnaires for cognitive performance during early child education. *Rev Chil Pediatr 88*(1), 35–40.

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.

A Appendix

Table A1: Summary of tests for quasi-random assignment: Analytical sample for age 1 socio-emotional outcomes

Outcome group	**# of outcomes**	**# of outcomes where mother Xs signficant**		
		No MHT	*Holm-Bonferroni-Sidak*	*Modified Romano-Wolf*
Nurse demographics	5	0	0	0
Nurse qualifications	12	0	0	0
Nurse experience	5	0	0	0
Nurse training	5	0	0	0
Process quality	10	3	1	0
Overall number	37	3	1	0
Overall %		8.1%	2.7%	0.0%

Table A2: Summary of tests for quasi-random assignment: Analytical sample for age 2 socio-emotional outcomes

Outcome group	**# of outcomes**	**# of outcomes where mother Xs signficant**		
		No MHT	*Holm-Bonferroni-Sidak*	*Modified Romano-Wolf*
Nurse demographics	5	0	0	0
Nurse qualifications	12	0	0	0
Nurse experience	5	0	0	0
Nurse training	5	0	0	0
Process quality	10	3	1	0
Overall number	37	3	1	0
Overall %		8.1%	2.7%	0.0%

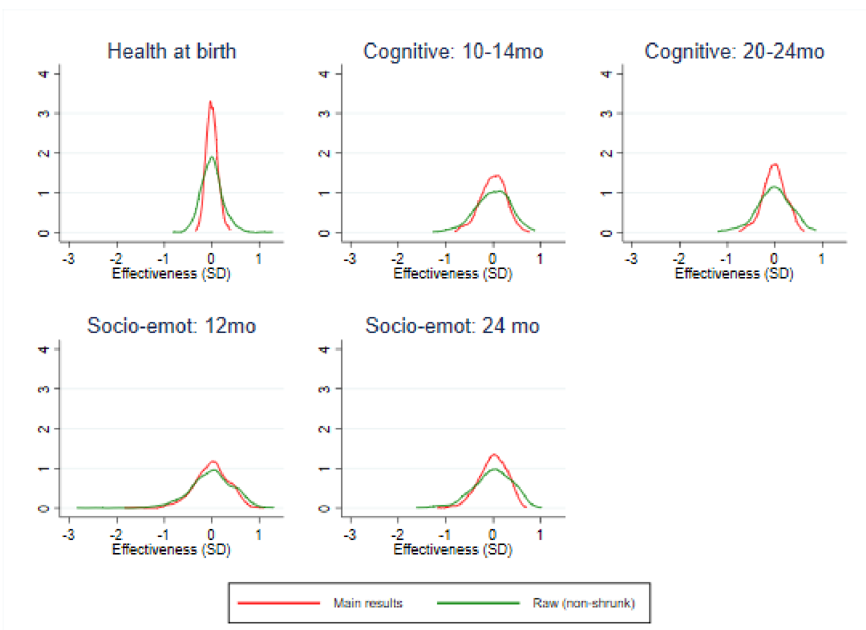
Table A3: Summary of tests for quasi-random assignment: Analytical sample for age 1 cognitive

Outcome group	**# of outcomes**	**# of outcomes where mother Xs signficant**		
		No MHT	*Holm-Bonferroni-Sidak*	*Modified Romano-Wolf*
Nurse demographics	5	0	0	0
Nurse qualifications	12	0	0	0
Nurse experience	5	0	0	0
Nurse training	5	0	0	0
Process quality	10	4	1	0
Overall number	37	4	1	0
Overall %		10.8%	2.7%	0.0%

Table A4: Summary of tests for quasi-random assignment: Analytical sample for age 2 cognitive outcomes

Outcome group	**# of outcomes**	**# of outcomes where mother Xs signficant**		
		No MHT	*Holm-Bonferroni-Sidak*	*Modified Romano-Wolf*
Nurse demographics	5	0	0	0
Nurse qualifications	12	0	0	0
Nurse experience	5	0	0	0
Nurse training	5	0	0	0
Process quality	10	3	1	0
Overall number	37	3	1	0
Overall %		8.1%	2.7%	0.0%

Figure A1: Impact of shrinkage on distribution of family nurse effectiveness



Note: See Figure 7. Main results have been adjusted using an empirical Bayes estimator ("shrinkage estimator"); raw estimates show the distribution before shrinkage is applied. Source: Authors' calculations using data provided by the Department for Health and Social Care.