Environment and Development Economics (2024), **29**, 359–378
doi:10.1017/S1355770X24000123

RESEARCH ARTICLE

The impact of water quality on children's education: evidence from 39 districts in the Ganges Basin of India

Md Ohiul Islam,[1]* and Moumita Ghorai[2]

[1]Economics Department, University of Nevada at Reno, Reno, NV, USA and [2]United Nations Development Programme, New York, NY, USA
*Corresponding author: Md Ohiul Islam; Email: oislam@unr.edu

(Submitted 5 August 2022; revised 30 March 2024; accepted 6 April 2024; first published online 16 May 2024)

Abstract
We investigate the effect of water quality on the educational outcomes of children aged 8–11 in 39 districts in five states in the Ganges Basin of India. Using data from the Centre for Pollution Control Board of India and the Indian Human Development Survey (IHDS) 2011–12, we study the effect of water quality in the Ganges Basin on the performance in three test scores. Our evidence suggests that faecal coliform levels in water sources above safety thresholds negatively affect reading and writing test scores. The effects of Nitrate-N and Nitrite-N in the water appear to be weaker compared to those of faecal coliform. The results establish that water pollution caused by excessive presence of faecal coliform is an important environmental factor in determining educational outcomes of children. High levels of faecal coliform in the water could be lowering cognitive abilities of the pollution-affected children through the channel of waterborne diseases.

Keywords: water pollution; water quality; pollution and education; cognitive abilities; children's education

JEL codes: I25; Q23; J24; J13

1. Introduction

The Ganges spans approximately 26 per cent of India's territory and sustains nearly half of its population (Chakraborti *et al.*, 2018). Despite its importance, it is becoming one of the world's most polluted rivers due to growing population, industrialisation, and urbanisation (Chaudhary and Walker, 2019). Urban areas near the Ganges saw a 30 per cent population increase from 2001 to 2011, which likely worsened the pollution (Government of India, 2011). Consequently, the pollution in the Ganges not only harms the environment but also poses significant health and economic consequences for the people living nearby (Das and Birol, 2010; Khan *et al.*, 2016; and others).

Many studies show that polluted water threatens public health and economic well-being. The Ganges, a key water source, is among the world's most polluted rivers

© The Author(s), 2024. Published by Cambridge University Press

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

(Chaudhary and Walker, 2019). Pollution can affect children's physical growth and cognitive development, as water filters may not remove all pollutants. This paper explores how pollution in the Ganges Basin affects the education of children aged 8–11 across 39 districts. Long-term exposure to pollution could impair cognitive abilities, potentially leading to lower educational achievements (Dewey *et al.*, 2023). We use data from the Central Pollution Control Board (2012*a*) and the 2011–12 wave of the Indian Human Development Survey (Desai and Vanneman, 2012) to analyse how organic and inorganic pollutants impact children's test scores. We focus on the effects of faecal coliform and Nitrate Nitrogen + Nitrite Nitrogen on children's reading, maths and writing abilities. For brevity, we will refer to Nitrate Nitrogen + Nitrite Nitrogen as Nitrate-N + Nitrite-N henceforth.

Originating from the Gangotri glacier in Uttarakhand, India, the Ganges flows 2,525 km across five states to the Bay of Bengal. It is essential for drinking, cooking and irrigation. However, pollution from sewage, industrial waste and agricultural runoff – exacerbated by population and industrial growth – poses a significant challenge. A recent report indicates that 764 industries release 500 million litres of wastewater into the Ganges daily.[1] Heavy metals in the water can cause kidney damage and cancer (Lellis *et al.*, 2019). Furthermore, long-term consumption of water with heavy metal content has been shown to impair cognitive function, according to several studies (Siegal and Share, 1990; Tolins *et al.*, 2014; Tyler and Allan, 2014). Nitrates and antibiotic-resistant bacteria in the water also pose health risks (Quist *et al.*, 2018; Adimalla, 2020). This study examines the impact of faecal coliform and Nitrate-N + Nitrite-N on children's cognitive abilities and educational outcomes, establishing an association between polluted water in the Ganges and lower test scores.

Religious activities such as ritual baths, idol immersion, and cremation add to the Ganges' pollution, increasing heavy-metal levels and the river's biochemical oxygen demand (BOD), often exceeding the Central Pollution Control Board (CPCB) standards. During the Maha Kumbh festival, studies of the Ganges water show that such mass gatherings significantly raise BOD, total suspended solids, and ammonia nitrogen beyond safe limits for outdoor bathing. The water also shows high levels of faecal and total coliforms, leading to more water-borne diseases (Tyagi *et al.*, 2013).[2]

Several studies have shown that the water quality of the Ganges is unsuitable for drinking and bathing at many monitoring points (Mariya *et al.*, 2019). This can pose a higher risk to human health (Chaudhri and Jha, 2012), and can potentially lead to lower cognitive abilities through the channel of health deterioration. When it comes to educational outcomes of children in the context of developing countries, researchers are more interested in socioeconomic and household conditions as determinants of children's education (Nambissan, 2009; Chaudhri and Jha, 2012). A growing literature provides evidence that exposure to pollutants, especially air pollutants, leads to lower educational outcomes in the US (Sanders, 2012; Rosofsky *et al.*, 2014; Ebenstein *et al.*, 2016; Roth, 2017). However, to the best of our knowledge, this is the first research that specifically examines the negative impact of poor water quality on educational outcomes in the context of a developing country like India.

[1] See ENVIS Centre on Control of Pollution Water, Air and Noise (CPCB) (2023) for more.

[2] Figures A1, A2 and A3 in the online appendix respectively depict Indian states the Ganges flows through, and the pollution intensity from faecal coliform and Nitrate-N + Nitrite-N in areas monitored by the CPCB, including the Ganges and its tributaries.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

This paper investigates the understudied area of pollution's impact on education in developing countries such as India. Water pollution leads to both immediate and long-term health issues, including negative effects on cognitive development from prolonged pollution exposure. Increased population density in polluted areas further exacerbates these effects, reducing children's cognitive abilities. Despite its importance, such research is limited, often overshadowed by urgent issues like child mortality. Moreover, while the discourse on environment and development prioritizes health and the reduction of child mortality, interest in educational outcomes often takes a backseat. Some studies that explored only the environmental and health outcomes were conducted after pollution control laws like the Ganga Action Plan were implemented (Dwivedi *et al.*, 2018). The lack of data for long-run health and cognitive outcomes is another hurdle in researching the connection between water pollution and children's educational outcomes.[3]

2. Data

To examine the relationship between the water quality of the river Ganges and children's educational outcomes, we merge two types of data: (1) household survey data, which provides information on children's educational outcomes, and (2) water quality data, encompassing various measures of water quality.[4] Below, we detail both data sources and describe the variables employed to estimate our empirical model.

2.1 Indian human development survey

The source of the household survey data for this paper is the Indian Human Development Survey (IHDS), a nationally representative dataset.[5] For this paper, we use the second round of the survey, conducted between November 2011 and October 2012. In this round, 42,152 households across 1,503 villages and 971 urban neighbourhoods throughout India were interviewed. While the first wave took place in the 2004–05 period, data from both the base year and the second round cannot be combined for this study because educational outcomes were only measured in the second round. Most children surveyed were at most two years old during the 2004–05 period and not suitable for educational aptitude testing. Data on various socioeconomic characteristics, such as individual health, household employment, and income, along with school facilities and staff, were collected. The interviews utilised two sets of questionnaires: one on income and social capital, typically answered by the male head of the household, and another

[3]Despite India's long history of environmental protection laws, such as the Water (Prevention and Control of Pollution) Act of 1974, the Air (Prevention and Control of Pollution) Act of 1981, and the Environment (Protection) Act of 1986, the country has continued to face challenges in enforcing pollution standards (Greenstone and Hanna, 2014). The CPCB and the State Pollution Control Boards (SPCBs) were established, and the government of India has adopted several environmental protection regulations over the past few decades. A landmark verdict by the Supreme Court in 1984, known as M.C. Mehta vs. Union of India, significantly reduced Ganges pollution and led to a decrease in neonatal mortality rate (Do *et al.*, 2018). This case marked the beginning of various initiatives aimed at cleaning the river. Following this, in 1985, the Ganga Action Plan was initiated to control water pollution in the Ganges, and it was subsequently expanded into the National River Conservation Plan, encompassing other rivers in India.

[4]District names serve as the common geographic identifiers between these two data sources.

[5]This dataset is made publicly available by Desai and Vanneman (2012). The IHDS is a biennial household panel survey.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

on education and health, answered by an ever-married woman. The collected data are organised into fourteen modules, of which the Individual, Household, and School Facilities modules are used for this study.[6] After merging the data and excluding missing values, we retain 1,147 observations for children aged 811 living in 39 districts across five states in the Ganges Basin, where water quality was monitored.[7]

2.2 Water quality data

We gathered water quality data for the districts in the Ganges basin for the years 2012 and 2013, drawing from the CPCB (2012*a*) database. This database operates under the Ministry of Environment, Forest, and Climate Change of the Indian Government.[8] The CPCB selects monitoring points along rivers or near water bodies (lakes and groundwater sources) that likely exhibit varying levels of key pollutants and potential turbidity. Monitoring points within districts along a river are sometimes categorised as either upstream or downstream from well-known locations. With each monitoring point's specific location provided, we identify the nearest district to each point. For instance, if a monitoring point is in a river, we assign it to the district situated directly on the riverbank. Most districts in our sample are located by a river, on the banks of the Ganges and/or Yamuna, or along their tributaries.[9]

Pollution data was collected quarterly and monthly at these monitoring points, with CPCB publishing yearly averages for minimum, mean and maximum levels of each water quality indicator. For example, at a specific monitoring point j at time $t = 1$, the CPCB calculates the minimum, mean and the maximum levels of faecal coliform, $F_{\max,1,j}$, $F_{\text{mean},1,j}$, and $F_{min,1,j}$, respectively. By averaging these measurements over total T periods, they create $(\sum_{t=1}^{T} F_{\max,t,j})/T$, $(\sum_{t=1}^{T} F_{\text{mean},t,j})/T$, and $(\sum_{t=1}^{T} F_{\min,t,j})/T$. If a district has J monitors – the monitor index being $j = 1, 2, 3, \ldots, J$ – and if data was collected by CPCB at T times in 2012, then we calculate the district mean of faecal coliform as $(\sum_{j}^{J} \sum_{t=1}^{T} F_{\text{mean},t,j})/(T \times J)$. We use this averaging scheme for each district. Compared to the average maximum and minimum levels of pollution exposure, represented by $(\sum_{j}^{J} \sum_{t=1}^{T} F_{\max,t,j})/(T \times J)$ and $(\sum_{j}^{J} \sum_{t=1}^{T} F_{\min,t,j})/(T \times J)$ respectively, the overall mean pollution level $(\sum_{j}^{J} \sum_{t=1}^{T} F_{\text{mean},t,j})/(T \times J)$ more accurately indicates the level of pollution to which the sample respondents were most frequently exposed. The minimum and maximum readings from the monitoring points may reflect infrequent dips and spikes in pollution, not necessarily representing the regular exposure levels for children. Since CPCB provides only minimum and maximum readings at each monitoring point but not their frequencies, we decide to use only the mean pollution levels from the monitoring points to calculate $(\sum_{j}^{J} \sum_{t=1}^{T} F_{\text{mean},t,j})/(T \times J)$, the district-level

[6]Individual, Household, Eligible Women, Birth History, Medical Staff, Medical Facilities, Non-Resident, School Staff, School Facilities, Wage and Salary, Tracking, Village, Village Panchayat, Village Respondent.

[7]The merged data comprise 204,575 household members from 42,152 households. Of these household members, 27,670 are under the age of 12. Maths, reading, and writing tests were administered to 11,749 individuals under 12. After removing around 100 missing values in control variables, we are left with 1,147 children in our analysis. These children reside in districts near the Ganges, Yamuna, or their tributaries within the Ganges Basin, where CPCB monitored various water sources.

[8]Total coliforms organism and faecal coliform are very similar indicators. We only use faecal coliform in this study.

[9]Five districts in our sample had the Ganges or Yamuna flowing through them. In two districts, Jhansi and Gaya, the CPCB monitored only groundwater and lake water.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

pollution measure. District-level means of the other water quality variables have been calculated in the same way.

We primarily use water quality data from 2012, supplementing it with 2013 data to fill any gaps. Missing readings for certain monitoring points in 2012 could potentially bias the computation of average water quality variables. To address this, we impute missing values using their 2013 counterparts. We found that readings from monitoring points available in both years were consistent, with no cases of monitors shifting from benign pollution levels in 2012 to hazardous levels in 2013. Therefore, we are confident that our approach to handling missing data ensures the reliability and representativeness of the actual pollution levels.

2.3 Descriptive statistics

Table 1 displays mean values for key variables, with each column representing a sample based on the type of water source monitored for pollution. For instance, the averages in the first column are derived from data on children in districts where river water was monitored. Column 7 in table 1 shows variable means for the full sample of 1,147 children. In some districts, more than one type of water source was monitored. According to columns 1 and 2 in table 1, mean faecal coliform and mean Nitrate-N + Nitrite-N levels are higher in the 'river' and 'Ganges' samples compared to 'Yamuna', 'groundwater' (GW) and 'Tributaries' (Trib.). The main binary variables of interest are district-average 1[Mean faecal Coliform $> 2,500$ MPN/100 ml] and 1[Mean Nitrate $- N +$ Nitrite $- N > 1$ mg/1 L]. For simplicity and to save space, we express these variables as $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$ using Iverson notation, respectively.[10]

Table 1 displays significant variations in the average values of water pollution measures. For example, the highest mean faecal coliform level is observed in the 'Lake' sample, while the 'Ganges' sample records the highest mean levels of Nitrate-N + Nitrite-N. Conversely, the 'Yamuna' sample, shown in column (3), has the lowest levels of both mean faecal coliform and mean Nitrate-N + Nitrite-N, coinciding with the lowest mean test scores. These patterns indicate a possible link between higher district test scores and elevated levels of pollutants, possibly because urban districts, despite higher pollution, often have access to better educational resources and means to counteract water pollution effects. Hence, the descriptive data in table 1 alone cannot comprehensively evaluate pollution's negative impact on test scores. A detailed analytical model is essential to pinpoint the impact of pollution exposure on test scores.

Our study focuses primarily on district-mean levels of faecal coliform and Nitrate-N + Nitrite-N as the main water pollutants, rather than on other pollutants for which data are available. Other water quality metrics, such as biochemical oxygen demand

[10]The bars over FCOLI and NIT denote that they represent means. The term 'limit' is used to indicate their respective safe levels. Both variables indicate if the respective pollution amounts are above individual acceptable limits. The Indian CPCB (2012*b*) sets the acceptable limit of faecal coliform at 2,500 MPN/100 ml, where MPN means 'most probable number'. Its limit is set at 2,500 MPN/100 ml by the Indian CPCB (2012*b*). They inspected whether, in 100 millilitres of water, the most probable count of coliform colonies was above 2,500. The data from the Indian CPCB (2012*a*) does not include a limit for NITRATE- N+ NITRITE-N (mg/l). The World Health Organisation (2011) provides separate safety limits for Nitrate-N and Nitrate-N, which are 10 mg/l and 1 mg/l, respectively. Using the 1 mg/l limit, the more restrictive of the two limits, we create the binary indicator $1[\overline{\text{NIT}} > \text{limit}]$; the value 1 indicates that the NITRATE-N + NITRITE-N level has exceeded 1 mg/1.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

Table 1. Analytical sample means of key variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables	River	Ganges	Yamuna	Lake	GW	Trib.	All
Mean faecal coliform (MPN/100 ml)[a]	2.27	2.44	0.06	6.25	1.69	1.77	1.15
Mean Nitrate-N/Nitrite-N (mg/l)	1.13	1.22	0.34	0.49	1.05	0.65	0.89
Mean biochemical oxygen demand (mg/l)	4.23	3.77	5.88	7.53	3.61	4.68	4.86
Mean dissolved oxygen (mg/l)	7.08	7.29	6.27	6.15	7.00	6.94	6.96
Mean pH	7.75	7.78	7.57	7.71	7.67	7.62	7.66
1[Faecal coliform > 2,500 MPN/100 ml]	0.84	0.83	1.00	0.84	0.68	0.59	0.72
1[Nitrate − N + Nitrite − N > 1 mg/l]	0.22	0.24	0.00	0.18	0.35	0.30	0.27
1[BOD >3 mg/l]	0.57	0.53	1.00	0.63	0.40	0.33	0.43
1[DO <4 mg/l]	0.25	0.19	0.49	0.44	0.21	0.23	0.23
1[pH <6.5 mg/l or pH >8.5 mg/l]	1.00	1.00	1.00	1.00	1.00	0.95	0.96
Reading test *Z*-score	0.16	0.21	−0.12	0.32	0.19	0.13	0.13
Maths test *Z*-score	0.19	0.23	−0.09	0.45	0.26	0.21	0.20
Writing test *Z*-score	0.17	0.22	−0.12	0.43	0.22	0.18	0.16
Age	9.51	9.53	9.51	9.59	9.52	9.49	9.48
Sex – 1 if Male	0.49	0.50	0.53	0.55	0.52	0.53	0.52
1 [Majority religious group]	0.52	0.53	0.64	0.56	0.48	0.51	0.53
Anthropometry – height (cm)	128.06	128.15	129.00	126.43	126.60	126.32	127.13
Anthropometry – weight (kilograms)	25.73	25.87	25.30	26.36	25.63	25.17	25.32
1[HH expenditure ⩽25th ptile][b]	0.23	0.25	0.27	0.17	0.27	0.27	0.25
1[HH expenditure ⩽50th ptile]	0.45	0.46	0.48	0.39	0.53	0.54	0.50
1[HH expenditure ⩽75th ptile]	0.69	0.70	0.70	0.68	0.77	0.88	0.75
School distance (kilometres)	1.56	1.57	1.66	1.99	1.56	1.53	1.57
School hours/week	30.73	30.73	33.64	27.7	29.40	29.44	30.13
Private tuition hours/week	3.86	4.06	1.23	5.33	5.12	4.62	4.11
Books uniform cost (thousand Rs.)	0.89	0.88	1.26	0.99	0.65	0.73	0.84
Short-term morbidity (days)	1.22	1.28	1.01	1.01	1.01	0.96	1.08
1[Water is purified in HH][c]	0.10	0.11	0.05	0.15	0.09	0.77	0.09
1[HH has indoor piped water supply]	0.15	0.16	0.07	0.23	0.11	0.90	0.11
1[HH has water drinking vessel]	0.71	0.69	0.76	0.76	0.70	0.68	0.71
1[Always handwash][d]	0.75	0.72	0.76	0.77	0.75	0.69	0.72
N	576	532	155	206	769	738	1,147

Notes: Columns (1) to (7) show variable means for district groups by water source type monitored in 2012. Columns (1) to (6) (detail specific sources: Ganges and Yamuna (1), only Ganges (2), only Yamuna (3), lakes (4), groundwater (5), and tributaries (6), with column (7) combining all districts.
[a]Mean faecal coliform (MPN/100 ml), reported in millions.
[b]HH expenditure: Household per capita expenditure.
[c]Household purifies water by boiling, filtering, aquaguard, or chemicals.
[d]Members of the households always wash hands after defaecation.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

(BOD), dissolved oxygen level (DO) and pH, are not classified as pollutants, though they do assess water quality.[11] We incorporate these metrics as control variables in our model. It is important to note that BOD and DO levels do not consistently correlate with the levels of our primary pollutants of interest. Typically, higher BOD levels and lower DO levels are observed in more turbid water, which may coincide with higher levels of faecal coliform and Nitrate-N + Nitrite-N (Ahipathy and Puttaiah, 2006). However, the absence of undesirable BOD and DO levels does not necessarily mean the absence of unsafe levels of faecal coliform and Nitrate-N + Nitrite-N. For instance, table 1 indicates that the groundwater sample exhibits relatively fewer occurrences of undesirable BOD and DO levels, yet the mean faecal coliform level in these districts is very similar to that of the full sample. In addition, in districts adjacent to the Yamuna River where BOD levels exceed preferred thresholds, Nitrate-N + Nitrite-N levels do not reach hazardous levels. Thus, BOD and DO levels do not always serve as accurate indicators of pollution. Lastly, the pH level exhibits minimal variation across the samples mentioned in columns 1 to 7 of table 1.[12] All these samples, along with almost all districts in 'tributaries' and the full sample, maintain high but safe pH levels. Consequently, overall pH levels do not present a significant risk to the cognitive abilities of children.

In table 1, individual characteristics such as age, gender, height, weight, and family consumption expenditure show only marginal variation across the monitored water source categories. Interestingly, the proportion of households with indoor piped water supply and those purifying water vary between 0.05–0.77 and 0.09–0.77, respectively. Handwashing after defecation is a critical preventive measure against many diseases (Curtis and Cairncross, 2003), and the proportion of households consistently practicing this varies narrowly from 0.69 to 0.77. Table A2 represents variable means for samples that are exposed to unsafe levels of faecal coliform and Nitrate-N + Nitrite-N. Table A3 includes means of additional variables we use as controls. Note that all tables whose numbers are preceded by 'A' appear in the online appendix, in which we provide explanations of the table contents below the tables as needed.

For regression analysis, we employ binary measures of the water pollutants, $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$. Using binary variables offers three distinct advantages. First, they enable a clear distinction between the districts experiencing unsafe pollution levels and those that do not, based on the established safety limits for pollutant concentrations. Second, understanding the estimated effect of the binary variables that signal unsafe pollution levels in districts does not rely on pollution changing by a certain amount; there was not much difference in pollution levels from 2012 to 2013. Also, minute fluctuations, like a one MPN increase in faecal coliform in 100 ml of water, are unlikely to make noticeable differences in test scores, making the estimated effect of the one-unit hard to interpret. Lastly, identification of the effects of pollutants in a regression model can be challenging at extremely high values of the pollution-measuring continuous variables. This complexity arises because districts with the most significant river pollution are often both densely populated and economically advanced. It is easier

[11] BOD indicates the oxygen consumed by microorganisms. When more microorganisms are present in the water, decomposing waste matter and propagating, dissolved oxygen levels decrease. Consequently, these two variables are highly correlated (Jouanneau *et al.*, 2014).

[12] pH measures the acidity or alkalinity of water, with the scale ranging from 0 to 14. Values below 7 are acidic, and values above 7 are alkaline. Water with very low or high pH may indicate chemical or heavy metal pollution (U.S. Geological Survey, 2019).

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

for such districts to insure themselves against high levels of pollution by establishing superior water filtration systems.

We examine the educational outcomes of children living in Ganges Basin districts, focusing on areas where water sources were monitored for pollution. The survey assessed children's reading, writing and arithmetic skills through tests administered to all eligible children aged 8–11 in each household. As indicated in table 1, the test scores are considered continuous variables, with a comprehensive description provided in table A4.[13] These tests, developed in collaboration with researchers from PRATHAM,[14] were pretested to ensure they were comparable across various languages. This method allows us to analyse the educational performance of school children in different states, accommodating the diverse languages used as mediums of instruction. Despite each Indian state having its unique school curriculum, PRATHAM's tests remain consistent across the board. The standardisation of test scores enables us to assess the impact of pollution exposure on children's average position within the test score distribution.

3. Empirical model

The empirical model examines the effect of water quality on test scores (equation (1)). The analytical sample contains unique children $i = 1, 2, 3 \ldots n$ living in $k = 1, 2, 3, \ldots, K$ districts,

$$Z_{ik} = \alpha_{ik} + \boldsymbol{W}'\Theta + \boldsymbol{X}'\Gamma + \chi_{\mathrm{k}} + \epsilon_{ik}, \tag{1}$$

where $\boldsymbol{W}$ is the vector of water quality variables and their values vary between districts, $\boldsymbol{X}$ is a vector of X_{ik} control variables, and χ_k are district dummy variables. We use the same right-hand-side variables for each test outcome, Z_{ik}. The main treatment variables, $1[\overline{\mathrm{FCOLI}} > \text{limit}]$ and $1[\overline{\mathrm{NIT}} > \text{limit}]$, vary only between districts and not within each district. Our baseline model uses random intercept regression. ϵ_{ik} is the individual-level error term and Z_{ik} indicates our set of dependent variables are nested within cluster k, with each district representing a separate cluster. Since $1[\overline{\mathrm{FCOLI}} > \text{limit}]$ and $1[\overline{\mathrm{NIT}} > \text{limit}]$ vary between districts, we can interpret the coefficient estimates of these two variables as the average decline in the children's position within the test score distribution due to exposure to district-level pollutants.[15] We include district-mean pH, and binary indicators of BOD and DO in the vector $\boldsymbol{W}$ from equation (1).[16]

[13]When adding more control variables to test the robustness of our primary estimates, treating scores as ordinal or binary variables causes convergence issues in multinomial logistic/probit model estimations. Similar to our method, studies by Chudgar and Quin (2012) and Singhal and Das (2019) also consider test scores as continuous in their OLS model estimations, indicating that this approach does not compromise the insights gained.

[14]The tests were available in multiple languages. PRATHAM is a non-governmental organisation that supports social science research.

[15]For example, let us assume that the estimated effect of $1[\overline{\mathrm{FCOLI}} > \text{limit}]$ is statistically significant at −0.015 on maths test scores, which means that living in a district with unsafe levels of faecal coliform in its water sources causes the district's children to experience, on average, a drop of 0.015 standard deviations in their maths test score distributional position, effectively moving them to the left by 0.015 standard deviations in the score distribution.

[16]Only results in tables 2, 3 and 4 include mean BOD. Including both BOD and DO measures in regression specifications results in these variables' coefficient estimates having ambiguous signs. To simplify, we include both mean BOD and mean DO only in tables 2, 3 and 4 to avoid confusion.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

The economic intuition behind applying the random-effects model is that the district-level errors are not necessarily affecting Z_{ik} through the variables of interest, $\boldsymbol{W}$. Communities within a district can invest in water treatment plants and water supply networks to insure against pollution. More affluent districts, often more urbanised, tend to pool resources to develop better public water supply networks to mitigate water pollution risks (Sarker *et al.*, 2021). Since water supply networks are monopolies requiring an initial fixed investment, and marginal cost of water supply to additional households is low, all the households in a district would have the same quality of water supply network available for them irrespective of individual household-level wealth and income. In other words, both rich and poor participate in the same water distribution network and are subject to similar levels of water quality. Thus, the unobserved heterogeneity due to a district's water supply characteristics of a district can be considered as random intercepts, $E(\boldsymbol{X}|\chi_k) = 0$, for the households and are not likely to drive or be driven by the household-level observed variables in $\boldsymbol{X}$. If $E(\boldsymbol{X}|\chi_k) \neq 0$, then we would need fixed-effects estimation of equation (1). Therefore, we model district-level exposure to water quality as random district-level effects.[17]

We prefer a random-effects model over one with district fixed effects because the fixed- effects model can introduce multicollinearity between the district-level dummy variables and the binary pollution variables. We run different tests to check if the random-effects model should be used instead of some alternative models. Diagnostic tests developed by Hausman (1978) and Schaffer and Stillman (2006) show that the random-effects model is preferred over the fixed-effects model.[18] Additionally, a test by Breusch and Pagan (1980) shows that the random-effects model is favoured over a simple ordinary least squares (OLS) model. Furthermore, we conduct a likelihood-ratio (LR) test that indicates that a random-effects model is preferred to a pooled model with district dummy controls. Overall, the results support applying a random intercept (district-level) specification.

The binary variables indicating unsafe levels of faecal coliform and Nitrate-N + Nitrite-N correlate with DO, BOD, and pH to some degree, as they all reflect aspects of water quality. The exact functional relationships between them are unknown. Generally, water quality deteriorates when faecal coliform and Nitrate-N + Nitrite-N exceed safety limits. Consequently, the estimated effect of main water pollution measures may be overstated, capturing both the overall water quality impact and specific pollution contents. However, water turbidity is also associated with poor quality, making it essential to control for the effects of mean BOD, mean pH and mean DO in equation (1). By doing so, we might have overly adjusted for water quality effects, rendering the estimates of the impact of unsafe levels of faecal coliform and Nitrate-N + Nitrite-N as 'lower-bound' estimates.

[17]In less-developed rural areas where (publicly funded) water supply networks are not established and water treatment plants are privately owned, the ability to insure against low water quality varies only at the community level, not at the household level. We control for the effect of this insurance ability using water-supply related controls in our model.

[18]Schaffer and Stillman (2006) provide a test for over-identifying restrictions in random-effects versus fixed-effects models. The fixed effects estimator relies on the orthogonality conditions that W_k, each variable in the $\boldsymbol{W}$ vector (equation (1)), is uncorrelated with the idiosyncratic error ε_{ik}, i.e., $E(W_k \times \varepsilon_{ik}) = 0$. The random effects estimator introduces additional orthogonality conditions that W_k are uncorrelated with the group-specific error χ_k (the 'random effects'), i.e., $E(W_k \times \chi_k) = 0$. These additional orthogonality conditions are over-identifying restrictions that we test. The results suggest considering a random-effects model.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

3.1 Identification

Equation (1) is based on the structure of a simple education production function. This function, widely discussed in the education economics literature, relates educational inputs to outcomes like test scores and class rankings (see Krueger (1999) and Hanushek (2010), among others). We assume that water quality levels are 'predetermined' factors in the education production process. Thus, the error term ε_{ik} is uncorrelated with water quality, or $E(\boldsymbol{W}|\varepsilon_{ik}) = 0$. While this is a strong assumption, we later introduce a propensity score matching model to estimate the causal effects of $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$ on test scores, relaxing this initial assumption.

River pollution is the outcome tied to economic activities, population density and geographic characteristics of an area. However, schooling is governed by state policies and government mandates in India, i.e., all children must attend schools (Chhokar, 2010). The government provides funding to the schools and dictates school curricula and related policies (Kingdon, 2007). The average quality of education and outreach at a district is not subject to the aggregate factors which may drive river pollution – overpopulation, urbanisation and industrialisation. Average education outcomes of the children may be driven by river pollution and other aggregate factors. Pollution impacts education production through the channel of both short-term and long-term health, as health is directly linked to water quality and, consequently, to productive outcomes such as educational attainment.

The CPCB employs stringent criteria to select monitoring points, indicating a non-random selection process. Consequently, the non-random selection of monitoring stations leads to a non-random selection of districts in our analysis. To address this, we calculate district-level mean pollution after aggregating readings from all monitoring points in a district. If the sample distribution of pollutants is skewed right because CPCB monitors more polluted areas, then the sample mean might exceed the true average pollution level. However, our focus is on binary indicators that show whether average monitored pollution levels exceed safety limits. Given that the sample includes districts with pollution levels below the unsafe threshold, it seems unlikely that CPCB exclusively monitored the most polluted river sections. Furthermore, some monitors detected no faecal coliform and Nitrate-N + Nitrite-N levels, suggesting that the selection of monitoring sites is unlikely to compromise the validity of our findings on the pollutants' treatment effect.

For robustness checks, the vector $\boldsymbol{X}$ in equation (1) is expanded to include the effects of teaching quality, educational expenditure, schooling quality, short-term morbidity, use of technology, and household members' personal hygiene. Since we lack variables for long-term morbidity throughout the children's lives, which could be linked to river pollution, we use district-level short-term morbidity as a proxy. The decline in skills such as maths, reading and writing cannot result from random sickness episodes alone. Short-term morbidity does not reveal the children's susceptibility to illness. Continuous consumption of poor-quality water, even if it does not cause immediate sickness, may lead to cognitive declines in children. The reading, writing and maths tests administered by Pratham (2021) measure the students' average cognitive abilities. Therefore, mean district-level morbidity is intended to capture spikes in short-term morbidity due to unforeseen reasons and the overall health of children in the district, excluding the cognitive loss channel in children exposed to unsafe pollution levels in drinking water.

We investigate the possible channels of cognitive ability loss due to pollutant contents in drinking water. Thus, we further demonstrate that interaction terms between $1[\overline{\text{FCOLI}} > \text{limit}]$ and binary variables describing household water supply and storage

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

choices are statistically significant. This analysis aims to identify how water pollutants not removed by the water supply system – which may or may not have a filtration system – affect children's cognitive abilities.[19]

Household characteristics such as the educational level of the head, available resources, and income significantly impact children's educational outcomes. Families with well-educated heads, ample resources, and higher incomes often see better educational results for their children. However, when considering the substantial impact of high water-pollution levels on education and income, children from households with lower educational outcomes may become trapped in a cycle of poverty. These children may face challenges in earning low incomes and lack the means to relocate from areas with poor water quality. In such a scenario, the current household head's lower investment in children's education might be linked to lower investment (P_k) in his/her education when he/she was a child and therefore, $E(P_k|\epsilon_{ik}) \neq 0$. In addition, the observational data used here does not include individual or household-level instruments that could be used to infer causation between poor water quality and educational outcomes.

We define a binary treatment variable T_f in the following way:

$$T_f \begin{cases} 1 & \text{if } \overline{\text{FCOLI}} > \text{limit} \\ 0 & \text{otherwise} \end{cases}.$$

Therefore, we estimate average treatment effect on the treated (ATT), which measures the difference between expected test scores of children in high-pollution districts $T_f = 1$ versus a counterfactual outcome expressed as:

$$\begin{aligned} \text{ATT}_f &= E[Z_1 - Z_0|T_f = 1] \\ &= E[Z_1|T_f = 1] - E[Z_0|T_f = 1]. \end{aligned} \tag{2}$$

In equation (2), Z_0 and Z_1 are outcomes of the non-treated ($T_f = 0$) and the treated ($T_f = 1$). The subscript f expresses that the treatment is unsafe levels of faecal coliform. $E[Z_0|T_f = 1]$ is the counterfactual state that we do not observe and estimate. By extension, the ATT is also applicable for unsafe levels of Nitrate-N + Nitrite-N. If T_n holds 1 for district-level mean Nitrate-N + Nitrite-N to be over the safe level, and 0 otherwise, then $\text{ATT}_n = E[Z_1 - Z_0|T_n = 1] = E[Z_1|T_n = 1] - E[Z_0|T_n = 1]$. The subscript n expresses that the treatment is unsafe levels of Nitrate-N + Nitrite-N. Identification is dependent on the assumption of conditional independence – if we control for the household and individual factors that drive educational outcomes, then the treatment effect can be considered random. For this non-experimental exercise, we use the widely known propensity score matching (PSM) developed by Rosenbaum and Rubin (1983).[20]

The baseline regression results in tables 2–4 can be combined to provide a picture of the negative impact of river pollution on children's test outcomes. Column 1 results are estimated using the full sample in each of the three tables. The pollutants do not appear to generate a statistically significant effect on the test scores which are based on the full

[19]While we account for the effects of district-level short-term morbidity, this channel could receive mixed effects from other externalities associated with river pollution. For instance, consuming fish from a polluted river could also impair children's cognitive functions in the long term (Singh and Soma, 2014). Another potential externality is the use of polluted water for irrigation, which might bypass the water supply system and affect health (Singh *et al.*, 2020).

[20]We implement PSM using the algorithm described in chapter 24 of Cameron and Trivedi (2022).

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

Table 2. Baseline regression – the effect of water pollution on reading test score

	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Reading	Reading	Reading	Reading	Reading
	Score	Score	Score	Score	Score	Score
AGE	0.0949	0.122	0.0980	0.0908	0.0913	0.120
	(0.0252)	(0.0597)	(0.0354)	(0.0253)	(0.0356)	(0.0318)
FEMALE	0.00724	0.0586	0.0229	0.0144	0.0138	0.0154
	(0.0479)	(0.0641)	(0.0856)	(0.0596)	(0.0796)	(0.0519)
HEIGHT	0.00275	0.00138	−0.00244	0.00117	−0.00141	0.000889
	(0.00272)	(0.00562)	(0.00428)	(0.00233)	(0.00448)	(0.00257)
WEIGHT	0.0140	0.00848	0.0198	0.0154	0.0221	0.0128
	(0.00514)	(0.00915)	(0.00613)	(0.00562)	(0.00638)	(0.00670)
HH con.≤75th ptile	−0.0125	0.216	−0.0561	0.0953	−0.0305	0.0147
	(0.0720)	(0.0963)	(0.114)	(0.111)	(0.103)	(0.103)
HH con.≤50th ptile	−0.0965	−0.357	−0.150	−0.116	−0.168	−0.0951
	(0.0673)	(0.139)	(0.0982)	(0.106)	(0.0930)	(0.132)
HH con.≤25th ptile	−0.278	−0.196	−0.294	−0.286	−0.198	−0.285
	(0.0680)	(0.136)	(0.0814)	(0.132)	(0.109)	(0.137)
Indoor piped water	0.225	0.101	0.244	0.241	0.261	0.288
	(0.0882)	(0.123)	(0.0955)	(0.114)	(0.0996)	(0.137)
$1[\overline{FCOLI} > limit]$	−0.129	−0.749	−0.234	−0.0689	−0.245	−0.0578
	(0.0933)	(0.313)	(0.115)	(0.134)	(0.124)	(0.0814)
$1[\overline{NIT} > limit]$	−0.0812	−0.0459	−0.0650	−0.193	−0.119	−0.140
	(0.104)	(0.103)	(0.170)	(0.0607)	(0.172)	(0.0999)
$1[\overline{DO} < threshold]$	−0.0752	−0.0169	0.131	−0.171	0.0688	−0.0318
	(0.0987)	(0.327)	(0.0896)	(0.203)	(0.138)	(0.165)
Mean BOD	0.00291	0.00647	0.0190	0.0312	0.00126	0.00130
	(0.00434)	(0.00925)	(0.00805)	(0.0233)	(0.0255)	(0.00375)
Mean pH	−0.160	−1.603	−0.171	−0.167	0.0684	−0.233
	(0.123)	(0.613)	(0.109)	(0.197)	(0.177)	(0.198)
N	1,147	206	532	769	576	738
Overall R^2	0.27	0.33	0.30	0.53	0.31	0.51
Sample	All	Lake	Ganges	GW	River	Trib.

HH con., Household consumption per capita; ptile, percentile; GW, groundwater; Trib., Tributaries.
Notes: Robust standard errors clustered at district level in parentheses.
Explanatory variables not reported: Numerical variables such as 'hours spent at school per week', 'hours spent doing homework per week', 'hours spent being tutored per week', 'distance from school to home', 'number of days the child spent disabled because of short-term morbidity in the last 30 days'. Binary variables such as '1 = Rupees spent on books and uniform > Rs. 500', '1 = water storage vessel available at home', '1 = water is purified at home though some mode of filtration or boiling', '1 = household members always wash hands after defaecation'.

sample. Only for the 'river' and the 'Ganges' samples do we see unsafe levels of faecal coliform generating a statistically significant negative impact.[21] The largest impact of faecal

[21] We remind the readers that the CPCB of India monitored groundwater and lakes in some districts. We consider all districts where any water source is monitored and which are in states through which the rivers Ganges, Yamuna, and their tributaries flow.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

Table 3. Baseline regression - the effect of water pollution on maths test score

	(1)	(2)	(3)	(4)	(5)	(6)
	Score	Score	Score	Score	Score	Score
AGE	0.0667	0.0231	0.0999	0.0689	0.0875	0.0653
	(0.0256)	(0.0411)	(0.0468)	(0.0325)	(0.0448)	(0.0241)
FEMALE	−0.0685	0.0543	−0.0294	−0.0553	−0.0306	−0.0458
	(0.0414)	(0.154)	(0.0783)	(0.0643)	(0.0753)	(0.0719)
HEIGHT	0.00394	0.0149	−0.000741	0.00267	0.000194	0.00473
	(0.00281)	(0.00295)	(0.00427)	(0.00353)	(0.00406)	(0.00329)
WEIGHT	0.0148	0.00479	0.0222	0.0124	0.0241	0.0130
	(0.00645)	(0.00700)	(0.00659)	(0.00884)	(0.00623)	(0.00891)
HH con.≤25th ptile	−0.259	−0.474	−0.370	−0.259	−0.273	−0.226
	(0.0898)	(0.203)	(0.0675)	(0.127)	(0.0988)	(0.131)
HH con.≤50th ptile	−0.0245	−0.0155	−0.0206	−0.0489	−0.0365	−0.0000922
	(0.0969)	(0.127)	(0.150)	(0.128)	(0.137)	(0.141)
HH con.≤75th ptile	−0.246	−0.115	−0.379	−0.250	−0.315	−0.238
	(0.0919)	(0.133)	(0.152)	(0.106)	(0.144)	(0.0976)
Indoor piped water	0.138	0.169	0.0510	0.158	0.0687	0.220
	(0.0896)	(0.227)	(0.140)	(0.106)	(0.133)	(0.0799)
1[$\overline{\text{FCOLI}}$ > limit]	−0.146	−0.669	−0.322	−0.0913	−0.342	−0.131
	(0.128)	(0.311)	(0.132)	(0.126)	(0.138)	(0.0911)
1[$\overline{\text{NIT}}$ > limit]	−0.0493	0.112	0.0868	0.0168	0.0282	−0.0912
	(0.160)	(0.175)	(0.107)	(0.114)	(0.111)	(0.127)
1[$\overline{\text{D.O.}}$ < threshold]	−0.0545	−0.0524	0.115	−0.137	0.0611	0.00650
	(0.163)	(0.357)	(0.167)	(0.230)	(0.156)	(0.189)
Mean BOD	0.000479	0.000237	0.0123	0.0176	−0.00489	−0.00351
	(0.00391)	(0.0117)	(0.0161)	(0.0237)	(0.0285)	(0.00374)
Mean pH	−0.372	−1.468	−0.335	−0.326	−0.111	−0.488
	(0.196)	(0.484)	(0.178)	(0.262)	(0.238)	(0.172)
N	1,147	206	532	769	576	738
Overall R^2	0.28	0.56	0.34	0.27	0.33	0.26
Sample	All	Lake	Ganges	GW	River	Trib.

HH con., Household consumption per capita; ptile, percentile; GW, groundwater; Trib., Tributaries.
Notes: Robust standard errors clustered at district level in parentheses.
Explanatory variables not reported: Numerical variables such as 'hours spent at school per week', 'hours spend doing homework per week', 'hours spent being tutored per week', 'distance from school to home', 'number of days the child spent disabled because of short-term morbidity in the last 30 days'. Binary variables such as '1 = Rupees spent on books and uniform > Rs. 500', '1 = water storage vessel at home', '1 = water is purified at home though some mode of filtration or boiling', '1 = household members always wash hands after defaecation'.

coliform is on the writing test and the smallest on the reading test when the samples, 'river' and the 'Ganges' are considered (columns 1 and 5 in tables 2–4). Overall, faecal coliform has a negative impact on test outcomes. Unsafe levels of Nitrate-N + Nitrite-N only has a significant impact on reading tests when 'groundwater' districts are considered. Among other variables, age, height, and weight have some estimated positive impact on the test scores as expected. Binary indicators of household consumption is coded 1 if per capita consumption expenditure of a household is at the 25th, 50th

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

Table 4. Baseline regression – the effect of water pollution on writing test score

	(1)	(2)	(3)	(4)	(5)	(6)
	Score	Score	Score	Score	Score	Score
AGE	0.0951	0.0429	0.114	0.106	0.103	0.128
	(0.0281)	(0.0541)	(0.0434)	(0.0313)	(0.0421)	(0.0374)
FEMALE	0.0536	0.140	0.0266	0.0717	0.00104	0.101
	(0.0348)	(0.0829)	(0.0607)	(0.0424)	(0.0609)	(0.0478)
HEIGHT	0.00206	0.00317	0.00217	0.00367	0.00342	0.00103
	(0.00278)	(0.00819)	(0.00399)	(0.00256)	(0.00408)	(0.00313)
WEIGHT	0.00759	0.0120	0.00544	0.00378	0.00820	0.00161
	(0.00507)	(0.00883)	(0.00724)	(0.00525)	(0.00775)	(0.00614)
HH con.≤25th ptile	−0.326	−0.586	−0.344	−0.280	−0.261	−0.288
	(0.0952)	(0.294)	(0.0985)	(0.123)	(0.123)	(0.127)
HH con.≤50th ptile	−0.0394	−0.0116	−0.168	−0.0335	−0.153	−0.0539
	(0.0613)	(0.190)	(0.0968)	(0.0912)	(0.0869)	(0.0858)
HH con.≤75th ptile	−0.0342	0.0235	−0.132	−0.0278	−0.0577	0.0335
	(0.0865)	(0.138)	(0.0736)	(0.104)	(0.0898)	(0.132)
Indoor piped water	0.168	−0.0467	0.0863	0.125	0.104	0.339
	(0.0888)	(0.140)	(0.110)	(0.114)	(0.103)	(0.126)
1[$\overline{\text{FCOLI}}$ > limit]	−0.170	−0.341	−0.351	0.00234	−0.364	−0.0119
	(0.110)	(0.326)	(0.178)	(0.136)	(0.183)	(0.154)
1[$\overline{\text{NIT}}$ > limit]	0.109	−0.0870	0.0851	0.00774	0.0307	0.0798
	(0.149)	(0.206)	(0.172)	(0.0932)	(0.186)	(0.136)
1[$\overline{\text{D.O.}}$ < threshold]	−0.137	−0.116	−0.0176	−0.218	−0.0700	−0.147
	(0.107)	(0.377)	(0.142)	(0.110)	(0.162)	(0.127)
Mean BOD	0.00236	0.0156	0.0183	0.0175	0.00139	0.000363
	(0.00280)	(0.0112)	(0.00758)	(0.0220)	(0.0249)	(0.00248)
Mean pH	−0.111	−0.627	−0.125	0.0458	0.108	−0.233
	(0.0948)	(0.520)	(0.138)	(0.174)	(0.177)	(0.155)
N	1,147	206	532	769	576	738
Overall R^2	0.20	0.31	0.26	0.31	0.44	0.23
Sample	All	Lake	Ganges	GW	River	Trib.

HH con., Household consumption per capita; ptile, percentile; GW, groundwater; Trib., Tributaries.
Notes: Robust standard errors clustered at district level in parentheses.
Explanatory variables not reported: Numerical variables such as 'hours spent at school per week', 'hours spend doing homework per week', 'hours spent being tutored per week', 'distance from school to home', 'number of days the child spent disabled because of short-term morbidity in the last 30 days'. Binary variables such as '1 = Rupees spent on books and uniform > Rs. 500', '1 = water storage vessel available at home', '1 = water is purified at home though some mode of filtration or boiling', '1 = household members always wash hands after defaecation'.

and 75th percentile of the distribution or below. As the reference group is children from households above the 75th percentile of the per capita consumption expenditure distribution, the estimated effects of these variables, when statistically significant, understandably are negative.

Having an indoor piped water supply is also estimated to have a positive impact on children's reading test scores (columns 1 and 3–6 in table 2), and also on maths and reading test scores (column 6 in tables 3 and 4). In districts adjacent to groundwater and

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

tributaries that were monitored for pollution, the effect of unsafe levels of faecal coliform and Nitrate-N + Nitrite-N are statistically indistinguishable from zero.[22] We investigate whether the interaction between unsafe levels of faecal coliform and access to indoor piped water supply significantly affects test scores. While indoor piped water alone has minimal impact on scores, column 6 in table A5 reveals that in the 'river' sample, the positive effect of indoor piped water (+0.818) on writing scores is nearly cancelled out by its interaction with the faecal coliform variable (−0.803). This suggests that faecal coliform may impair children's cognitive abilities, as reflected in test scores, despite the presence of indoor piped water supply. The results in columns 3 and 5 in table 3 are based on 'Ganges' and 'groundwater' samples. Tables 2–4 support the impact of unsafe levels of faecal coliform being primarily driven by the pollution in the river Ganges. Our other binary variable of interest about Nitrate-N + Nitrite-N only has a significant impact on reading test scores when the districts where groundwater is monitored are chosen.

We look for heterogeneity in the estimated effect of $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$ between genders. Looking for differential pollution effect on boys versus girls, we find that $1[\overline{\text{FCOLI}} > \text{limit}]$ has approximately 0.01 standard deviation greater effect on boys than girls in writing tests (columns 9 and 12 in table A6).[23]

Caste-based and religion-based discrimination in accessing safe water suggests that water pollution's impact might vary across different castes and religious groups (Hoff, 2016). However, dividing the sample by religion and caste results in too few observations per group, leading mostly to inconclusive results and hindering our ability to detect potential heterogeneity in the effects of $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$. Given the distinct social statuses and relationships among the six religious and caste groups, merging these groups to enlarge sample sizes could lead to misleading conclusions.

In table 5, we present ATT by estimating a PSM model as outlined in equation (2). The estimated ATT shows causal impact of the main pollution treatments. The results show that when the full samples are considered, T_f has a statistically significant causal impact on reading, maths and writing scores. T_n also has a negative impact on reading and maths scores.

3.2 Robustness checks

We check the robustness of the effects of the pollutants in several ways. We check if the effects $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$ differ across states. We find that the more economically developed West Bengal sees greater negative impact of $1[\overline{\text{FCOLI}} > \text{limit}]$ on writing tests compared to the Uttar Pradesh and Bihar-Jharkhand sample (columns 6 and 9 in table A10).[24] Next, we include more variables in $\boldsymbol{X}'\Gamma$ (equation (1)) that cover more factors related to individual characteristics, household characteristics, water source information, short-term morbidity and schooling. The results in tables A11 show if the effects of $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$ on reading and writing are robust even after the inclusion of a long list of control variables. The results in table A12 are estimated by adding indicators related to teaching quality to the regression specification in

[22] We cannot provide estimates separately for the districts adjacent to the river Yamuna where its water was tested for pollution because Nitrate-N + Nitrite-N and faecal coliform have no variation for those districts.

[23] Table A7 shows the male-female mean test score differences. Tables A7, A8 and A9 show results from attempts to tease out channels that could negatively affect female test scores.

[24] In this table, the coefficient for $1[\overline{\text{NIT}} > \text{limit}]$ for the sample of West Bengal is not identified as it has no variations in that state.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

Table 5. Average treatment effect on the treated

	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Maths	Writing	Reading	Maths	Writing
	Score	Score	Score	Score	Score	Score
T_f (1 versus 0)	−0.0882 (0.0412)	−0.265 (0.0762)	−0.143 (0.0120)			
T_n (1 versus 0)				−0.314 (0.0648)	−0.256 (0.0825)	−0.119 (0.0844)
N	1,147	1,147	1,147	1,147	1,147	1,147

Notes: Abadie and Imbens (2016) robust standard errors in parentheses. $T_f = 1$ means that the household is in district that received the treatment of exposure to unsafe levels of faecal coliform and $T_f = 0$ means untreated. $T_n = 1$ means that the household is in district that received the treatment of exposure to unsafe levels of Nitrate-N + Nitrite-N and $T_n = 0$ means untreated. Average treatment effect on the treated has been estimated by propensity-score matching. We consider a logit treatment model. Conditioning variables in the treatment model: demographic identities, age, height, weight, consumption expenditure by households, and individual-level variables: household per capita income, school distance, school hours/week, homework hours/week, private tuition hours/week, expenditure on books and uniform, short-term morbidity (days of disability in the previous thirty days before the survey interview), Binary: whether the household boils water for purification (1 = yes), whether household members wash hands after defaecation (1 = yes).

addition to the set of explanatory variables corresponding to the results in table A11.[25] The estimated effect of $1[\overline{\text{FCOLI}} > \text{limit}]$ on reading and writing scores is still robust in table A12.

As a sensitivity analysis, we estimate the baseline results using mixed-model specifications where the random-effects are interpreted as district-specific random intercepts (table A13). The estimated effect of $1[\overline{\text{FCOLI}} > \text{limit}]$ in table A13 are similar to those in tables 2–4, proving that these alternative specifications do not change the baseline results. In addition, tables A14 and A15 exhibit the statistically robust effects of $1[\overline{\text{FCOLI}} > \text{limit}]$ and $1[\overline{\text{NIT}} > \text{limit}]$, respectively employing two-level and three-level random-intercept models that account for variations within villages, neighbourhoods and households. In table A16, we find that after including a measure of short-term morbidity, the effects of $1[\overline{\text{FCOLI}} > \text{limit}]$ on reading scores in the 'river' sample and $1[\overline{\text{NIT}} > \text{limit}]$ on reading scores in the full sample remain robust statistically. Next, after adding state-specific controls to our regression specifications, we find that the effect of $1[\overline{\text{NIT}} > \text{limit}]$ loses its statistical significance but the effect of $1[\overline{\text{FCOLI}} > \text{limit}]$ remains statistically robust on the three test scores for the full sample (table A17). We attempt to separate the seasonality effect from the pollution effect in table A18. As our dataset is of a cross-sectional nature, we plug State ID × Disctrict mean morbidity × Survey month – interaction terms – into the model, which are supposed to account for variations in district-mean morbidity over the survey months, and find that the effect of $1[\overline{\text{FCOLI}} > \text{limit}]$remains robust on reading and maths scores in the full-sample regression (table A18).

Besides water pollution, other types of pollution like land and air pollution may also affect test scores. An increase in water and air pollution when both are driven by rapid urbanisation can coincide, and the estimated effect of water pollutants can partially

[25]Description of these variables can be found in table A3.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

contain the effect of air pollution. We have included PM2.5,[26] a measure of air pollution, as a control variable in our model. PM2.5 refers to particulate matter in the air that are less than 2.5 micrometres in diameter. We find that the impact of $1[\overline{\text{FCOLI}} > \text{limit}]$ on reading and maths scores remains statistically significant in the full sample in table A19. Moreover, its influence on writing scores also proved to be statistically significant in districts near Ganges. Our final robustness checking strategy instruments the district-mean level of faecal coliform with the district's upstream adjacent district's mean level of faecal coliform (MeanFCOLI). This instrumentation is based on the idea that pollution from an upstream district generates exogenous variation in its downstream neighbouring district; the upstream district is not likely to be influenced by downstream conditions. The effect of instrumented MeanFCOLI on reading scores across three different samples – full sample, 'river', and 'tributaries' sample – are reported in table A20.

The section 'Explanation for Table A20' in the online appendix includes the instrumentation strategy. We also observe weaker effect of the instrumented MeanFCOLI on the maths score in the full sample and the 'tributaries' sample but not on the writing score, potentially due to a smaller number of observations available. Notably, in the 'tributaries' sample, the coefficients for district MeanFCOLI remain unchanged between the random-effects and generalised 2SLS random-effects model (columns 13 to 18 in table A20). This instrumental variable analysis, leveraging upstream faecal coliform levels, acts as an additional robustness check, supporting our primary findings.

4. Conclusion

This study focuses on the impact of water pollution on the educational outcomes of school-going children aged 8–11 across 39 districts in the Ganges Basin of India. Water, as a crucial natural resource for production and consumption, can have long-term effects on human health, life expectancy, and cognitive functions through various channels. Using data from the CPCB of India and the IHDS 2011–12, we estimate water pollution's effect on performance in three tests taken by children aged 8–11 as part of the IHDS. We find that unsafe faecal coliform levels have a consistently robust negative effect on reading and writing test scores. In several extended specifications and sensitivity analyses, the impact of faecal coliform on maths scores was not statistically robust. The negative effect of Nitrate-N + Nitrite-N was statistically indistinguishable from zero in some robustness checks. The negative effects of faecal coliform in water sources on children's reading and writing performance prove to be consistently significant, even when controlling for additional factors such as average district-level short-term morbidity in children (over thirty days), quality of teaching, and adjustments made using a PSM model. This suggests that faecal coliform contamination may impair the cognitive development of children exposed to poor water quality through the channel of health deterioration for prolonged periods (exceeding 30 days). Future studies employing larger datasets and more precisely pinpointed water pollution data have the potential to refine our understanding of

[26] PM2.5, also known as particulate matter 2.5, refers to tiny airborne particles with a diameter of 2.5 microns or less. These particles are commonly measured in micro-grams per cubic meter ($\mu g/m^3$) to determine their concentration in the air. To ensure a safe and healthy environment, health regulations and standards are established to control and restrict the levels of PM2.5 present in the atmosphere. Presently, the WHO guidelines advocate for an annual average PM2.5 concentration of 5 micro-grams per cubic meter ($\mu g/m^3$) of air (World Health Organisation, 2021). Our records encompass the yearly mean PM2.5 data at the district level. We collect this data from the Energy Policy Institute at the University of Chicago (2023).

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

how water contaminants like faecal coliform and Nitrate-N + Nitrite-N impact cognitive functions.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1355770X24000123.

Acknowledgements. The views expressed in this research paper are those of the authors and do not reflect the positions or opinions of any affiliated organizations.

Data. Data from the Indian Human Development Survey are openly available at Inter-university Consortium for Political and Social Research, Ann Arbor, Michigan, The United States (URL: https://doi.org/10.3886/ICPSR36151.v6).

The water quality data is derived from the publicly available source at URL: https://cpcb.nic.in/nwmp-data-2012. This data is collected and maintained by Central Pollution Control Board (CPCB), Ministry of Environment, Forests and Climate Change, Government of India.

Financial support. We did not receive any financial support for the research, authorship, and/or publication of this article.

Competing interests. The authors declare none.

References

Abadie A and Imbens GW (2016) Matching on the estimated propensity score. *Econometrica* **84**, 781–807.

Adimalla N (2020) Spatial distribution, exposure, and potential health risk assessment from nitrate in drinking water from semi-arid region of south India. *Human and Ecological Risk Assessment: An International Journal* **26**, 310–334.

Ahipathy M and Puttaiah E (2006) Ecological characteristics of Vrishabhavathy river in Bangalore (India). *Environmental Geology* **49**, 1217–1222.

Breusch TS and Pagan AR (1980) The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies* **47**, 239–253.

Cameron AC and Trivedi PK (2022) *Microeconometrics Using Stata: Volume II: Nonlinear Models and Causal Inference Methods*. College Station, TX: Stata Press, StataCorp LLC.

Central Pollution Control Board (CPCB) (2012*a*) Water quality data 2012. Available at https://cpcb.nic.in/nwmp-data-2012/

Central Pollution Control Board (CPCB) (2012*b*) Annual report. Available at https://cpcb.nic.in/annual-report.php

Chakraborti D, Singh SK, Rahman MM, Dutta RN, Mukherjee SC, Pati S and Kar PB (2018) Groundwater arsenic contamination in the Ganga river basin: a future health danger. *International Journal of Environmental Research and Public Health* **15**, 180.

Chaudhary M, Walker T (2019) River Ganga pollution: causes and failed management plans (correspondence on Dwivedi *et al.* 2018. Ganga water pollution: a potential health threat to inhabitants of Ganga basin. *Environment International* 117, 327–338). *Environment International* **126**, 202–206.

Chaudhri DP and Jha R (2012) Child poverty and compulsory elementary education in India: policy insights from household data analysis. *Indian Journal of Human Development* **6**, 5–30.

Chhokar KB (2010) Higher education and curriculum innovation for sustainable development in India. *International Journal of Sustainability in Higher Education* **11**, 141–152.

Chudgar A and Quin E (2012) Relationship between private schooling and achievement: results from rural and urban India. *Economics of Education Review* **31**, 376–390.

Curtis V and Cairncross S (2003) Effect of washing hands with soap on diarrhoea risk in the community: a systematic review. *The Lancet Infectious Diseases* **3**, 275–281.

Das S and Birol E (2010) Estimating the value of improved wastewater treatment: the case of river Ganga, India. *Journal of Environmental Management* **91**, 2163–2171.

Desai S and Vanneman R (2012) India Human Development Survey-II (IHDS-II), 2011–12. Inter-university Consortium for Political and Social Research [distributor], 2018–08–08. Available at https://doi.org/10.3886/ICPSR36151.v6

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

Dewey D, England-Mason G, Ntanda H, Deane AJ, Jain M, Barnieh N, Giesbrecht GF and Letourneau N (2023) Fluoride exposure during pregnancy from a community water supply is associated with executive function in preschool children: a prospective ecological cohort study. *Science of The Total Environment* **891**, 164322.

Do QT, Joshi S and Stolper S (2018) Can environmental policy reduce infant mortality? Evidence from the Ganga pollution cases. *Journal of Development Economics* **133**, 306–325.

Dwivedi S, Mishra S and Tripathi RD (2018) Ganga water pollution: a potential health threat to inhabitants of Ganga basin. *Environment International* **117**, 327–338.

Ebenstein A, Lavy V and Roth S (2016) The long-run economic consequences of high-stakes examinations: evidence from transitory variation in pollution. *American Economic Journal: Applied Economics* **8**, 36–65.

Energy Policy Institute at the University of Chicago (2023) Air quality life index database. Available at https://aqli.epic.uchicago.edu/the-index/

ENVIS Centre on Control of Pollution Water, Air and Noise (CPCB) (2023) Annual progress report (2016–2017) Available at https://cpcbenvis.nic.in/annual_report_main.html

Government of India (2011) Census of India 2011. Available at https://censusindia.gov.in/census.website/data/census-tables

Greenstone M and Hanna R (2014) Environmental regulations, air and water pollution, and infant mortality in India. *American Economic Review* **104**, 3038–3072.

Hanushek EA (2010) Education production functions: evidence from developed countries. *Economics of Education* **2**, 132–136.

Hausman JA (1978) Specification tests in econometrics. *Econometrica: Journal of the Econometric Society* **46**, 1251–1271.

Hoff K (2016) Caste system. World Bank policy research working paper 7929. Washington, DC: World Bank.

Jouanneau S, Recoules L, Durand M, Boukabache A, Picot V, Primault Y, Lakel A, Sengelin M, Barillon B and Thouand G (2014) Methods for assessing biochemical oxygen demand (BOD): a review. *Water Research* **49**, 62–82.

Khan MYA, Gani KM and Chakrapani GJ (2016) Assessment of surface water quality and its spatial variation: a case study of Ramganga river, ganga basin, India. *Arabian Journal of Geosciences* **9**, 28.

Kingdon GG (2007) The progress of school education in India. *Oxford Review of Economic Policy* **23**, 168–195.

Krueger AB (1999) Experimental estimates of education production functions. *The Quarterly Journal of Economics* **114**, 497–532.

Lellis B, Fávaro-Polonio CZ, Pamphile JA and Polonio JC (2019) Effects of textile dyes on health and the environment and bioremediation potential of living organisms. *Biotechnology Research and Innovation* **3**, 275–290.

Mariya A, Kumar C, Masood M and Kumar N (2019) The pristine nature of river Ganges: its qualitative deterioration and suggestive restoration strategies. *Environmental Monitoring and Assessment* **191**, 1–33.

Nambissan GB (2009) *Exclusion and discrimination in schools: Experiences of Dalit children*. Indian Institute of Dalit Studies and UNICEF.

Pratham (2021) Pratham. Available at http://www.pratham.org

Quist AJ, Inoue-Choi M, Weyer PJ, Anderson KE, Cantor KP, Krasner S, Freeman LEB, Ward MH and Jones RR (2018) Ingested nitrate and nitrite, disinfection by-products, and pancreatic cancer risk in postmenopausal women. *International Journal of Cancer* **142**, 251–261.

Rosenbaum PR and Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosofsky A, Lucier CA, London B, Scharber H, Borges-Mendez R and Shandra J (2014) Environmental ascription in Worcester County, MA: toxic pollution and education outcomes. *Local Environment* **19**, 283–299.

Roth S (2017) Air pollution, educational achievements, and human capital formation. IZA World of Labor 2017, IZA Institute of Labor Economics, Bonn, Germany.

Sanders NJ (2012) What doesn't kill you makes you weaker: prenatal pollution exposure and educational outcomes. *Journal of Human Resources* **47**, 826–850.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123

Sarker B, Keya KN, Mahir FI, Nahiun KM, Shahida S and Khan RA (2021) Surface and ground water pollution: causes and effects of urbanization and industrialization in South Asia. *Scientific Review* **7**, 32–41.

Schaffer ME and Stillman S (2006) Xtoverid: Stata module to calculate tests of overidentifying restrictions after xtreg, xtivreg, xtivreg2, xthtaylor. Available at https://econpapers.repec.org/software/bocbocode/s456779.htm

Siegal M and Share DL (1990) Contamination sensitivity in young children. *Developmental Psychology* **26**, 455.

Singh AK and Soma G (2014) Assessment of human health risk for heavy metals in fish and shrimp collected from Subarnarekha River, India. *International Journal of Environmental Health Research* **24**, 429–449.

Singh V, Nagpoore NK, Chand J and Lehri A (2020) Monitoring and assessment of pollution load in surface water of river ganga around Kanpur, India: a study for suitability of this water for different uses. *Environmental Technology & Innovation* **18**, 100676.

Singhal K and Das U (2019) Revisiting the role of private schooling on children's learning outcomes: evidence from rural India. *South Asia Economic Journal* **20**, 274–302.

Tolins M, Ruchirawat M and Landrigan P (2014) The developmental neurotoxicity of arsenic: cognitive and behavioural consequences of early life exposure. *Annals of Global Health* **80**, 303–314.

Tyagi VK, Bhatia A, Gaur RZ, Khan AA, Ali M, Khursheed A, Kazmi AA and Lo SL (2013) Impairment in water quality of Ganges River and consequential health risks on account of mass ritualistic bathing. *Desalination and Water Treatment* **51**, 2121–2129.

Tyler CR and Allan AM (2014) The effects of arsenic exposure on neurological and cognitive dysfunction in human and rodent studies: a review. *Current Environmental Health Reports* **1**, 132–147.

U.S. Geological Survey (2019) pH and water. Available at https://www.usgs.gov/special-topics/water-science-school/science/ph-and-water

World Health Organisation (2011) Hardness in drinking-water background document for development of WHO guidelines for drinking-water quality. Available at https://cdn.who.int/media/docs/default-source/wash-documents/wash-chemicals/hardness-bd.pdf?sfvrsna13853a9_4

World Health Organisation (2021) WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Available at https://www.who.int/publications/i/item/9789240034228

Cite this article: Islam MO, Ghorai M (2024). The impact of water quality on children's education: evidence from 39 districts in the Ganges Basin of India. *Environment and Development Economics* **29**, 359–378. https://doi.org/10.1017/S1355770X24000123

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 15 Mar 2025 at 17:54:44, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1355770X24000123