Health Knowledge, Caste and Social Networks in India[†]

Niels-Hugo Blunch
Washington and Lee University & IZA
blunchn@wlu.edu

Nabanita Datta Gupta
Aarhus University & IZA
ndg@asb.dk

This Version: September 27, 2013

JEL Classifications: I12, I14, I15, O15

Keywords: Health knowledge, caste, social networks, discrimination, human capital, instrumental variables, matching, causal effects, India.

Abstract:
Addressing several methodological shortcomings of the previous literature, this paper explores the relationship among health knowledge and caste and a number of important mediating factors in India--attempting at estimating causal impacts through a combination of instrumental variables and matching methods, where possible. The results indicate the presence of a substantively large health knowledge caste gap (favoring high caste women) and also provides evidence that while observed individual characteristics such as education, information exposure, and access to social networks explain part of the gaps, a substantial part of the health knowledge gap is left unexplained. Overall, these results are consistent with the presence of discrimination towards low caste women in terms of health knowledge but at the same time also point towards the importance of continued attention towards education, institutions and economic policy for decreasing the health knowledge caste gap in India.

[†] We thank Ritwik Banerjee, Tor Eriksson, Art Goldsmith, Martin Paldam, and seminar participants at Aarhus University and Washington and Lee University for helpful comments and suggestions. Remaining errors and omissions are our own. The data were kindly provided by the Inter-university Consortium for Political and Social Research, Ann Arbor, MI, on behalf of Sonalde Desai, Reeve Vanneman, and the National Council of Applied Economic Research, New Delhi. The findings and interpretations are those of the authors and should not be attributed to any of the aforementioned individuals or organizations, however.

1. Introduction

Improved health conditions, arguably, are at the heart of human development. In order to improve health outcomes, however, people must first realize the availability and usefulness of health behaviors—that is, they must improve their health knowledge.

Addressing several methodological shortcomings of the previous literature, this paper addresses five questions regarding health knowledge and caste in India: (1) Does a ("raw") health knowledge gap exist between high and low castes in India?; (2) What happens to the caste gap in health knowledge as explanatory variables are added?; (3) What are the returns to education and information exposure in terms of health knowledge?; (4) What are the returns to social networks access in terms of health knowledge?; (5) What explains the observed caste gap in health knowledge—and is there evidence that this may be due to discrimination against the low caste?

The remainder of this paper is structured as follows. The next section motivates studying the inter-linkages among caste, social networks, and health knowledge in India. This is followed by a description of the data examined here in section three, while section four discusses the empirical strategy and related issues. Section five presents the results for the raw caste gap in health knowledge, as well as multivariate OLS and IV/2SLS regressions of the determinants of health knowledge and decompositions of the caste gap in both endowments and returns and in observables and unobservables—with all decompositions pursued both at the aggregate level, as well as tracing the determinants to their component parts, such as education, information exposure, and social network access. Finally, section six concludes and discusses policy implications and provides suggestions for further research.

2. Caste in India and Conceptual Framework

Indian society has been stratified according to an elaborate and rigid system of occupational specialization for thousands of years. The system of caste has resulted in the practice of extreme forms of prejudice against and the complete exclusion of certain groups from the opportunities for advancement. Caste discrimination is constitutionally illegal in India and starting in 1950, several government schemes have been implemented to improve the labor market and social conditions of the low caste. Yet, poor economic outcomes for the low caste persist. There are 5 major caste groups, which, over time have branched out into around 3000 subcastes or *jatis*. The most marginalized groups are the Scheduled castes (the "untouchables" or Dalits) and the Scheduled tribes (Adivasis, or indigenous peoples). Other Backward Castes (OBC) are placed higher in the hierarchy but still suffer economic disadvantage. Caste is endogamous and rarely can be changed. It is generally revealed by

the family name, village location, occupation or even dress and custom. Since caste is mainly associated with Hinduism, conversion to another religion or name changing is in principle a way to escape from one's caste of birth. The act of conversion or caste-free names, however, signals a prior low caste affiliation, and caste segmentation has even found its way into the other religions of India.

Why should caste be important for health knowledge? First of all, low caste individuals are disproportionately poor and have lower levels of education. In 2004/5, poverty rates among Dalits and Adivasis were 46% and 37% compared to 23% among non-SCs/STs (Chin and Prakash, 2011). Even though the educational attainment of the low caste has improved since the turn of the century, it lags behind that of upper caste Hindus (Borooah and Iyer, 2005). According to Grossman's 1972 model for the demand for health and health care, educated individuals both have greater allocative efficiency and greater productive efficiency when it comes to making investments in their health. Kenkel (1991) empirically shows that the more educated possessed greater levels of health knowledge and thereby enjoy an informational advantage.

Because of extreme prejudicial notions of contamination and loss of purity when encountering the low caste, the high-caste has traditionally forced them to live on the outskirts of the villages. This means lower proximity to health and educational facilities and thereby less contact with professional staff from whom health knowledge is obtained. By virtue of their isolation, they are exposed to less information. Even when they do establish contact, extreme discrimination excludes them from receiving the right treatment and knowledge.

A further consequence of low education is that the social networks of the low caste are of poorer quality—they would have greater *social distance* to members of the medical and educational communities, especially those in higher positions such as GPs and teachers and principals simply because too few members of the low caste are represented in these positions.

3. Data

The 2004/05 India Human Development Survey (IHDS) is a comprehensive nationally representative multi-purpose household survey of 41,554 households in 1,503 villages and 971 urban neighborhoods across India, and also collected information on communities. The IHDS survey was produced by the National Council of Applied Economic Research (NCAER), New Delhi, and the University of Maryland and used a multi-stage clustered sampling design, ensuring national representativeness of the

survey findings.[1] Relevant for the purpose here, for the subsample of ever-married women 15-49 years of age, the survey collected information on health information regarding a wide array of health issues—spanning areas as diverse as the correct treatment of diarrhea in children regarding their water intake to the "safe" period regarding the menstruation cycle—as well as education in of individual household members. The survey also collected household information such as age, information exposure, and access to social networks from a knowledgeable informant, typically the head of the household.

Since the dependent variable is health knowledge, the sample was first conditioned on ever-married females 15-49 years of age (36,130 observation). Additionally, since the institution of caste is much more prevalent in rural areas, the sample was then conditioned on ever-married females 15-49 years of age from rural areas, only (21,310 observations). To further enable focusing explicitly on caste, only Hindus and Tribals were kept in the sample; that is, Christians, Muslims and Sikh/Jain were excluded (16,986 observations). Lastly, some variables were missing for some women, so that the final estimation sample consists of 16,468 ever-married women from rural areas.

The dependent variable is the individual woman's health knowledge as measured across six different dimensions, based on the following questions: (1) Is it harmful to drink 1-2 glasses of milk every day during pregnancy?; (2) Do men become physically weak even months after sterilization?; (3) Do you think that the first thin milk that comes out after a baby is born is good for the baby, harmful for the baby, or it doesn't matter?; (4) Is smoke from a wood/dung burning traditional chulha good for health, harmful for health, or do you think it doesn't really matter?; (5) When children have diarrhea, do you think that they should be given less to drink than usual, more drink than usual, about the same, or it doesn't matter?; and, finally, (6) In which part of the menstrual cycle is a woman LEAST likely to get pregnant? The health knowledge measures for the dependent variables are then constructed as binary variables for answering correctly. Additionally, we construct a simple index summarizing a woman's overall health knowledge by summing across all six variables.

Explanatory variables include birth cohorts (constructed as a set of dummy variables spanning five years each), and a set of dummy variables for the highest level of education completed—spanning no education (the reference) through tertiary. Information exposure is constructed as two dummies based on the questions "How often do people in your household read the newspaper?" and "How often do people in your household watch TV?" with possible responses including "Never," "Sometimes," and "Regularly." Two dummies are then constructed for "Regular" usage, motivated

[1] See Desai et al (2010) for more details.

by the fact that that can be seen as the higher threshold.[2] The social networks variables examined here span two dimensions or sectors, namely health and education, and several levels: any knowledge to a person from within the sector, knowledge of particular levels of professionals within the sector, and finally whether or not this person is of the same jati as the respondent. We construct a series of dummy variables for any knowledge, high level knowledge (doctor for health, teacher or principal for education), and for whether the person is of the same jati as the respondent—thus leading to a total of six dummy variables. Additional variables in this analysis include access to health facilities in the community—specified as a set of dummy variables for availability of a Health Sub-center, Primary Health Center, Community Health Center, Government Maternity Center, Government Communicable Disease Facility (e.g., tuberculosis)—and district fixed effects. The means and standard deviations for the final estimation samples by caste are reported in Table A1, Appendix A.

4. Estimation Strategy and Related Issues

The conceptual framework discussed in Section 2 suggests that caste, educational attainment, information exposure, and access to social networks can directly affect the acquisition of health knowledge and also suggest additional factors that are potentially important for experiencing a teenage pregnancy and therefore should be included in the empirical specifications.[3] The empirical analysis will examine this relationship, using linear approximations of the health knowledge equation. The natural starting point is estimating the following regression by OLS[4] (i.e. as a Linear Probability Model, LPM):

$$HK_i = \alpha_0 + \alpha_1 CASTE + \alpha_2 EDU_i + \alpha_3 INFEXP_i + \alpha_4 SOCNET_i + \alpha_2 CONTROLS_i + \varepsilon_i, \qquad (1)$$

Where HK_i is either one of the six alternative binary health knowledge measures or the compositie (score) health knowledge index, $CASTE_i$ is a dummy variable for high caste, EDU_i is a set of dummies for educational attainment; $INFEXP_i$ is a vector containing the two dummy variables for information (newspaper and television); $SOCNET_i$ is a vector of social network access variables; and $CONTROLS_i$

[2] These are asked both for women and men in the household overall. We use the information pertaining to the women of the household here.

[3] At a minimum, if these factors are not included, one may systematically over- or underestimate the strength of the caste-health knowledge relationship.

[4] As is well known, there may be some concern about using OLS, or, in effect, the linear probability model (LPM), when the dependent variable is binary. For example, predicted probabilities may fall outside the (0,1)-range and heteroskedasticity also is present by default. However, it can be argued that the LPM approximates the response probability well, especially if (1) the main purpose is to estimate the partial effect of a given regressor on the response probability, averaged across the distribution of the other regressors, (2) most of the regressors are discrete and take on only a few values and/or (3) heteroskedasticity-robust standard errors are used in place of regular standard errors (Wooldridge, 2010). All three factors seem to work in favor of the LPM for the purposes of the application here.

is a vector of all additional controls, including district fixed-effects; and ε_i is an error-term capturing unobservables.

After thus presenting the benchmark estimation method, several issues needs to be addressed pertaining to the estimation of the caste-health knowledge relationship in equation (1)—where the most important arguably is the possible endogeneity of the education, information exposure, and social network access variables. Again, endogeneity has three possible causes: omitted variables, simultaneity, and measurement error, all of which are potentially relevant in this application. Regarding omitted variables, ability and preferences, for example, are unobserved and at the same time also main determinants of educational attainment, information exposure, and social network access. As a result, the estimated impacts of these variables may be affected by omitted variables bias. Second, simultaneity may be a potential issue, since obtaining health knowledge, information exposure, as well as access to social network access all involve choices on the part of the woman. Regarding measurement error, one important issue is that the variables for social network access are binary measures of access per se, that is, they do not measure the intensity of an individual's network access. Information exposure is also measured as binary variables and are also self-reported, both of which leads to measurement error.

One widely applied approach to deal with endogeneity involves instrumental variables (IVs), by estimating by Two-Stage Least Squares. It is often a daunting task, however, to come up with variables that are both highly correlated with the potentially endogenous variable(s) and which at the same time may also validly be excluded from the main equation. Arguably, human capital accumulation and skills acquisition depend on the availability of educational institutions, as well as their quality. This has led researchers to follow two main IV strategies in recent years: either using as IVs (1) various combinations of time of year, birth cohort, and/or geographical area of birth dummies to capture variation in institutional factors relevant for human capital accumulations such as compulsory schooling laws or expansion of educational programs (Angrist and Krueger, 1991; Duflo, 2001) or (2) variables for proximity or exposure to educational institutions in the local area (Card, 2001).

Since we don't have available the geographical area of birth, we first explored the second of these approaches for the case of educational attainment. It turned out, however, that the instruments were quite weak, thus leading to the so-called weak instruments problem (Staiger and Stock, 1997). Additionally, it can be argued that education is at least pre-determined, thus at least addressing the simultaneity-part of the potential endogeneity issue.

Turning to information exposure and social network access, where the endogeneity issues

seems particularly worrisome due to the likely string simultaneity of these variables vis-à-vis the dependent variable, namely health knowledge, one promising candidate is the share of the population in the area with regular information exposure and social network access (across the different sources of information and types/levels of social networks). The intuition behind this instrument—inspired by Gruber (2005)—is that the more information exposure or social network access there is in an area, the more likely it is that a given woman will be exposed to information from the media and/or gain access to social networks.

These considerations lead us to estimate the health knowledge equation with instrumental variables using the first stage equation:

$$X_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 CONTROLS_i + \nu_i, \quad (2)$$

where X_i is a binary, possibly endogenous, variable (information exposure and social network access), Z_i is a vector of instrumental variables, and $CONTROLS_i$ is a vector of all additional controls from the second stage regression (primarily included for efficiency), including all other (exogenous) variables. ν_i is an error-term capturing unobservables. The first-stage test for weak instruments then is performed as joint test on the variables in Z_i (the identifying instruments – that is, excluded from the second-stage regression). The second stage equation (the estimating equation) then includes the predicted values of the potentially endogenous variables from the first stage:

$$HK_i = \beta_0 + \beta_1 \hat{END}_i + \beta_2 CONTROLS_i + \gamma_i, \quad (3)$$

where HK_i measures the health knowledge of the ith woman, using either one of the six binary measures or the composite index measure; $\hat{END}_i$ is a vector of the fitted values of the potentially endogenous variables from the first-stage equation (2); $CONTROLS_i$ is a vector of all additional (exogenous) controls; and γ_i is an error-term capturing unobservables. Again, since (3) includes predicted variables as regressors, the standard errors must be adjusted accordingly. Further, so as to allow for arbitrary heteroskedasticity, the estimations of (1)-(3) will be carried out using Huber-White standard errors (Huber, 1967; White, 1980). To allow for the possibility that observations are correlated within communities the standard errors are also adjusted for within-cluster correlation (Wooldridge, 2010).

To help strengthen the social network results we also use matching on observables to estimate the impact of having any network access (only—since this estimation method only allows one "treatment." This estimation method goes straight towards constructing the proper counterfactual; that is, what would have happened in the absence of the treatment? Specifically, treatment and control groups are matched on observables and then the treatment effect is estimated as the mean difference

between the average effects between the matched samples (Rosenbaum and Rubin, 1983, 1984, 1985). "Treatment" for the application here refers to having any network access—so that we pursue two analyses in turn: either any health network access *or* any education network access, respectively. The counterfactual is approximated by the experiences of a "comparison group" of women who are similar in all respects except social network access. This is achieved using matching, which in practice amounts to using a two-stage approach. In the first stage participants and non-participants are matched, based on their observable characteristics. In the second stage the impact estimate—which corresponds to the estimate of α_1 or α_2 in (1) or β_1 in (3) from the regression case, depending on which "treatment" is considered—is then calculated as the difference in means of health knowledge outcomes between matched participants and non-participants.

There are several different ways to conduct the matching in practice. A simple and widely used method is "nearest neighbor" propensity score matching, where the match with the participant (a woman that has network access) is based on the closest non-participant (a woman which has no network access) in terms of the distances of their propensity score (the predicated probability of having social network access). One final issue related to matching is that the estimated treatment effect is only defined in the so-called "region of common support," which basically implies that the treatment and control groups must overlap in terms of their covariate values. To ensure this, we impose common support by excluding participant observations whose propensity scores are higher than the maximum or less than the minimum covariate values of the comparison group (as also suggested by Rosenbaum and Rubin, 1983).

After estimating the relationship between health knowledge and its main determinants as expressed by (1) and (3) above, the next step is to decompose the health knowledge gap into its main components using the Blinder-Oaxaca approach (Blinder, 1973; Oaxaca, 1973). The starting point of this approach is an OLS (or, in our case, an IV) regression of the outcome in question, estimated separately across the two relevant groups; here, high and low caste women, respectively (suppressing subscripts for individual women):

$$Y_H = \beta_H X + \varepsilon_H \quad (4)$$

$$Y_L = \beta_L X + \varepsilon_F \quad (5)$$

where Y_H and Y_L are health knowledge of high caste and low caste women, respectively; X is a vector of womens' characteristics (education, information exposure, social network access); β_H and β_L are the returns to the womens' characteristics; and ε_H and ε_L are error terms.

These caste stratified health knowledge regressions formally are merely inputs into the decomposition analysis. Specifically, the decomposition analysis amounts to examining to which extent the observed health knowledge gaps across caste are attributable to differences in the observable characteristics, to differences in the returns to those characteristics, and to the interaction of the two ("three-fold decomposition," see below for details) and, relatedly, to which extent the observed health knowledge gaps are due to observable and unobservable characteristics ("two-fold decomposition," see below for details). This analysis will comprise the second part of the multivariate empirical analysis and will be pursued as an Oaxaca-Blinder type decomposition.

Formally, following the methodology of Oaxaca (1973) and Blinder (1973), the difference in mean health knowledge for high and low caste women, denoted R, can be decomposed into three parts (Jann, 2008) using the empirical counterparts of equations (4) and (5) above:[5]

$$R = \bar{Y}_H - \bar{Y}_L = (\bar{X}_H - \bar{X}_L)\,\hat{\beta}_H + \bar{X}_H\,(H_M - \hat{\beta}_L) - (\bar{X}_H - \bar{X}_L)\,(\hat{\beta}_H - \hat{\beta}_L) \qquad (6)$$

This is a three-fold decomposition (Winsborough and Dickinson, 1971), where the first term represents the "endowments effect" and explains the differences that are due to individual characteristics (such as education, information exposure, social network access, etc). The second term reflects the "coefficients effect," which shows the differences in the estimated returns to high and low caste women's characteristics. Lastly, the third term, the "interaction effect," accounts for the fact that differences in endowments and coefficients between high and low caste women exist simultaneously. If high and low caste women obtain equal returns for their characteristics, the second and the third parts in equation (6) will equal zero and health knowledge differentials between high and low caste women will be explained by the differences in endowments alone.

The above decomposition is formulated based on the prevailing health knowledge structure of high caste women, i.e. the differences in endowments and coefficients between high and low caste women are weighted by the coefficients (returns) of high caste women. This seems reasonable for the application here, since the high caste dominates in the Indian society—and, thus, can be perceived as the non-discriminated-against group—as also revealed by the existence of substantial "raw" health

[5] In the following, bars on top of variables denote mean values, while $\hat{\beta}$ denotes estimated coefficient values from equations (1) and (2) above.

knowledge gaps presented in Table 1. This is therefore also the approach pursued in the subsequent analysis.[6]

An alternative approach, prominent in the literature on wage discrimination, is based on the assumption that wage differentials are explained by a unifying "non-discriminatory" coefficients vector, denoted β^*, which is estimated in a regression that pools together both of the two groups under consideration (here, high and low caste women). Then, the health knowledge gap can be expressed as:

$$R = \bar{Y}_H - \bar{Y}_L = (\bar{X}_H - \bar{X}_L)\,\hat{\beta}^* + \bar{X}_H(\hat{\beta}_H - \hat{\beta}^*) + \bar{X}_L\,(\hat{\beta}^* - \hat{\beta}_L) \qquad (7)$$

The above equation represents the so-called two-fold[7] decomposition:

$$R = Q + U \qquad (8)$$

Where $Q = (\bar{X}_H - \bar{X}_L)\,\hat{\beta}^*$ is the part of the health knowledge differential that is "explained" by sample differences assessed with common "returns" across the two groups and the second term $U = \bar{X}_H(\hat{\beta}_H - \hat{\beta}^*) + \bar{X}_L\,(\hat{\beta}^* - \hat{\beta}_L)$ is the "unexplained" part not attributed to observed differences in high and low caste characteristics. The latter part is often treated as discrimination in the literatures on gender and racial earnings gaps. It is important to note, however, that the "unexplained" part also captures all potential effects of differences in unobserved variables (Jann, 2008). And, to be sure, in the application here it is indeed possible to talk about "discrimination," per se, as being a low caste woman is an intrinsic characteristic. Again choosing the high caste health knowledge structure as the reference, (7) reduces to:

$$R = \bar{Y}_H - \bar{Y}_L = (\bar{X}_H - \bar{X}_L)\,\hat{\beta}_H + \bar{X}_L\,(\hat{\beta}_H - \hat{\beta}_L) \qquad (9)$$

Again, while the main analysis here takes the high caste health knowledge structure as the reference, several different specifications for the baseline specification (also known as the "absence of discrimination" specification), i.e. $\hat{\beta}^*$ in (7), will be pursued in the sensitivity analysis as a robustness check.

[6] Alternatively, however, this equation could also be represented based on the prevailing health knowledge structure of low caste women; this will be explored further in the sensitivity analysis.

[7] See Oaxaca (1973), Blinder (1973), Cotton (1988), Reimers (1983), Neumark (1988), and Jann (2008) for different approaches—basically, these differ in the relative weights they attribute to the two groups in the decomposition.

The standard errors of the individual components in equations (6) and (7) above are computed using the Delta method by applying the procedure detailed in Jann (2008), which extends the earlier method developed in Oaxaca and Ransom (1998) to deal with stochastic regressors.

In addition to examining the overall composition of the established health knowledge gaps, it would seem instructive to perform detailed decompositions, as well, whereby it is possible to see which explanatory variables contribute the most to the three- and/or two-fold overall decompositions. An issue here is that while the overall decompositions are always identified, the results for categorical variables in detailed decompositions depend on the choice of the reference category (Oaxaca and Ransom 1999). A possible solution to this problem is to apply the deviation contrast transformation to the estimates before conducting the decomposition (Yun 2005); this is also the approach pursued here. Similar to the OLS regressions, the decomposition estimations also all allow for arbitrary heteroskedasticity (Huber, 1967; White, 1980). So as to condense the wealth of results obtained here—thereby easing the interpretation of the many results—the detailed decompositions are done groupwise, rather than for each individual variable (for example, for education as a whole, rather than separately for by educational level, and so on). Here, too, the focus will be on the case where the high caste structure is taken as the reference, though the sensitivity analysis again will consider alternative specifications, as well.

5. Results

This section reviews the main results. This is centered on addressing the following five questions in turn:

(1) Does a ("raw") health knowledge gap exist between high and low castes in India?

(2) What happens to the caste gap in health knowledge as explanatory variables are added?

(3) What are the returns to education and information exposure in terms of health knowledge?

(4) What are the returns to social networks access in terms of health knowledge?

(5) What explains the observed caste gap in health knowledge—and, relatedly, is there evidence that this gap may be due to discrimination against the low caste?

It should be noted that since some of the tables are rather large, they have been placed in the Appendices (but are referred to, and discussed, in the body text below—and the pertinent excerpts of these tables are also presented in the relevant sections below).

Question 1: *Does a ("raw") health knowledge gap exist between high and low castes in India?*
From Table 1 below a raw health knowledge gaps is found across all six dimensions of health knowledge. While the estimated gap differs in magnitude—ranging from 1.8 percentage-points for the goodness of smoke from wood/dung burning to 11.4 percentage-points for the correct treatment of diarrhea in children, regarding their water intake—it is substantively large in most cases. We again note particularly the gap related to the correct treatment of diarrhea in children, regarding their water intake, since this should be of particular policy concern—being effectively a matter of life or death for these children. For the composite measure, high caste women answer on average about 0.35 questions more correctly than low caste women—since the average number of correctly answered questions of the latter is about three questions, this reflects a substantively large difference in the total combined average health knowledge across high and low caste women.

Table 1. Raw Health Knowledge Gaps Using Six Different Measures of Health Knowledge and Combined (Score) Index Measure of Health Knowledge

	(1) Milk drinking during pregnancy	(2) Physical weakness of men after sterilization	(3) Goodness of the first (thin) milk for the baby	(4) Goodness of smoke from wood/dung burning	(5) Treatment of diarrhea in children, re water intake	(6) Menstruation, re "safe period"	(7) Combined (score-) index
High caste	0.769***	0.334***	0.756***	0.798***	0.604***	0.154***	3.415***
	[0.010]	[0.011]	[0.010]	[0.009]	[0.012]	[0.008]	[0.029]
Low caste	0.712***	0.261***	0.697***	0.781***	0.490***	0.128***	3.068***
	[0.006]	[0.006]	[0.006]	[0.005]	[0.007]	[0.004]	[0.016]
Difference	0.057***	0.073***	0.059***	0.018*	0.114***	0.026***	0.347***
	[0.012]	[0.012]	[0.012]	[0.011]	[0.013]	[0.009]	[0.033]
N	16,468	16,468	16,468	16,468	16,468	16,468	16,468

Notes: Values in brackets are robust Huber-White (Huber, 1967; White, 1980) standard errors. ***: statistically significant at 1 percent; *: statistically significant at 10 percent.
Source: 2004/05 India Human Development Survey (IHDS).

Question 2: *What happens to the caste gap in health knowledge as explanatory variables are added?*
To start answering the second question, we first need to determine the preferred estimation method—where the main options are OLS/LPM) and 2SLS/IV. The results from specification tests indicate that

the use of 2SLS/IV seems warranted for this application overall (Table A2 in the Appendix). First, the results from Wu-Hausman tests indicate that information exposure and social network access are endogenous. Second, the results from the F-tests of the joint significance of the identifying instruments from the first stage of the 2SLS procedure indicate that the identifying instruments are highly correlated with all the potentially endogenous variables—with statistical significance levels of 1 percent or better in almost all cases, and with extremely high F-statistics, too.[8] It therefore seems prudent to use 2SLS/IV, since this is empirically relevant and will, thus, also address the endogeneity concerns discussed earlier. Nevertheless, we will present OLS results alongside those of 2SLS/IV as benchmarks, since the previous literature overwhelmingly has used OLS.

The overwhelming impression from Table 2 is that the caste gap narrows substantially when all explanatory variables (discussed in Section 3) are included in the regressions—for both the

Table 2. Caste Coefficient: Caste, only (Raw Gap) and for Full Specifications (OLS/LPM and 2SLS/IV)

	Milk 1	Sterilization	Milk 2	Smoke	Diarrhea	Menstruation	Score Index
Raw gap:							
High caste	0.057***	0.073***	0.059***	0.018	0.114***	0.026**	0.347***
	[0.016]	[0.017]	[0.016]	[0.017]	[0.020]	[0.013]	[0.049]
OLS/LPM, Full Specification:							
High caste	0.035***	0.027*	0.016	-0.019++	0.028**	-0.002	0.084**
	[0.013]	[0.014]	[0.014]	[0.012]	[0.014]	[0.010]	[0.036]
IV/2SLS, Full Specification:							
High caste	0.041***	0.030**	0.008	-0.020++	0.015	-0.011	0.063*
	[0.014]	[0.014]	[0.015]	[0.013]	[0.015]	[0.011]	[0.038]
N							

Notes: Robust Huber-White (Huber, 1967; White, 1980) standard errors, adjusted for within-cluster correlation/clustering (Wooldridge, 2010), in brackets under parameter estimates. ***: statistically significant at 1 percent; **: statistically significant at 5 percent; *: statistically significant at 10 percent; ++: statistically significant at 15 percent +: statistically significant at 20 percent. The full specifications include district fixed effects, age cohort dummies, and dummies for educational attainment, information exposure, and access to health facilities in the community.
Source: 2004/05 India Human Development Survey (IHDS).

[8] It should be noted that since the first-stage regression is exactly identified, Hansen's (1982) J-test for over-identification is not relevant for this application.

(benchmark) OLS regressions and for the (preferred) 2SLS/IV regressions, and across all health knowledge measures. In a few cases (goodness of the first (thin) milk of the baby and menstruation, regarding the "safe" period) the magnitude of the remaining caste gap shrinks so much that it looses statistical significance, as well as becoming practically nil in substantive terms. This indicates that personal characteristics ("endowments") are important for explaining the established raw caste health knowledge gaps—and since from the descriptive statistics (Table A1, Appendix A) high caste women are favored over low caste women for favorable characteristics such as education, information exposure, and social network access, this already hints at differences in characteristics/endowments becoming important determinants in the subsequent Oaxaca decompositions.

Question 3: *What are the returns to education and information exposure in terms of health knowledge?* In Section 2, we suggested that education and information exposure were among the main determinants of health knowledge. The evidence from Table 3 confirms this—for both OLS and 2SLS/IV. Across both estimation methods and all health knowledge measures (except knowledge of "safe" periods regarding menstruation) education is strongly associated with health knowledge—and the higher the education, the higher the health knowledge. For diarrhea, for example, completing primary education is associated with about 4 percentage-points higher probability of knowing the correct treatment, while this probability increases to about 9 percentage-points for completion of higher secondary (IV/2SLS). While there are some differences across the OLS and 2SLS/IV results for educational attainment, the estimated coefficients for information exposure are substantially higher for 2SLS/IV than for OLS—in line, for example, with the findings from the returns to education studies in the labor markets literature when education is endogenized (e.g., Card, 2001).

Table 3. Education and Information Exposure Coefficients: Full Specification (All Explanatory Variables, Including All Networks Variables)

	Milk 1	Sterilization	Milk 2	Smoke	Diarrhea	Menstruation	Score Index
OLS/LPM:							
Some education	0.013	0.022+	0.057***	0.037**	0.037**	-0.011	0.154***
	[0.017]	[0.016]	[0.017]	[0.014]	[0.017]	[0.011]	[0.042]
Primary	0.026**	0.058***	0.074***	0.027**	0.069***	0.018*	0.272***
	[0.013]	[0.013]	[0.014]	[0.011]	[0.015]	[0.010]	[0.036]
Middle/some sec	0.025++	0.128***	0.110***	0.057***	0.107***	0.009	0.434***
	[0.016]	[0.017]	[0.015]	[0.014]	[0.017]	[0.011]	[0.044]
Higher secondary	0.056**	0.220***	0.149***	0.079***	0.148***	0.007	0.659***
	[0.024]	[0.035]	[0.024]	[0.024]	[0.040]	[0.022]	[0.087]
Tertiary	0.046++	0.320***	0.166***	0.083***	0.139***	0.042	0.797***
	[0.030]	[0.040]	[0.024]	[0.026]	[0.041]	[0.047]	[0.083]
Reads newsp reg	0.004	-0.002	0.018	-0.005	0.040*	0.052**	0.107++
	[0.024]	[0.028]	[0.022]	[0.025]	[0.024]	[0.021]	[0.068]
Watches tv reg	0.035***	0.036***	0.004	0.038***	0.022*	-0.006	0.128***
	[0.013]	[0.012]	[0.012]	[0.010]	[0.013]	[0.008]	[0.032]
R^2	0.253	0.226	0.223	0.305	0.292	0.231	0.313
IV/2SLS:							
Some education	0.013	0.021	0.051***	0.031*	0.018	-0.009	0.125**
	[0.017]	[0.017]	[0.018]	[0.016]	[0.019]	[0.012]	[0.049]
Primary	0.028*	0.050***	0.062***	0.020++	0.042**	0.015+	0.217***
	[0.016]	[0.016]	[0.017]	[0.014]	[0.018]	[0.011]	[0.050]
Middle/some sec	0.023	0.116***	0.088***	0.044**	0.055**	-0.008	0.319***
	[0.026]	[0.026]	[0.024]	[0.021]	[0.026]	[0.016]	[0.075]
Higher secondary	0.060++	0.206***	0.122***	0.070**	0.091*	-0.023	0.526***
	[0.037]	[0.046]	[0.038]	[0.033]	[0.049]	[0.029]	[0.130]
Tertiary	0.028	0.296***	0.128***	0.063++	0.062	-0.001	0.576***
	[0.048]	[0.056]	[0.045]	[0.043]	[0.057]	[0.051]	[0.149]
Reads newsp reg	0.132	0.076	0.017	-0.023	0	0.165**	0.367
	[0.125]	[0.110]	[0.104]	[0.099]	[0.107]	[0.074]	[0.404]
Watches tv reg	0.053+	0.041	0.068*	0.074**	0.142***	0.037	0.415***
	[0.039]	[0.038]	[0.039]	[0.037]	[0.041]	[0.029]	[0.114]
N	16,468	16,468	16,468	16,468	16,468	16,468	16,468

Notes: Robust Huber-White (Huber, 1967; White, 1980) standard errors, adjusted for within-cluster correlation/clustering (Wooldridge, 2010), in brackets under parameter estimates. ***: statistically significant at 1 percent; **: statistically significant at 5 percent; *: statistically significant at 10 percent; ++: statistically significant at 15 percent +: statistically significant at 20 percent. All specifications include district fixed effects. Additional controls include a dummy for high caste, age cohort dummies, and dummies for educational attainment, information exposure, and access to health facilities in the community.
Source: 2004/05 India Human Development Survey (IHDS).

Question 4: *What are the returns to social networks access in terms of health knowledge?*

To examine the importance of access to social networks for health knowledge, due to the likely presence of strong multicollinearity among the network access variables, we first add only access to any health network to the core model (all explanatory variables except social network access variables: Table 4); then access to any education network (Table 5); and finally all network variables (Table 6). From Table 4, there appears to be a fairly strong relationship (both in statistical and substantive terms) between having access to any health network and at least some of the health knowledge measures, as well as the composite measure of overall health knowledge. In particular, the measure indicating correct treatment of diarrhea in children regarding their water intake exerts a strong association with access to any health network: about 5 percentage-points for OLS and about 20 percentage-points for IV/2SLS. Perhaps surprisingly there appears to be negative and statistically significant relationship between Adding access to any education network mostly does not change results—except for OLS, where there appears to be an additional, separate effect, which is both substantively and statistically significant in a few cases. The main impression when adding all additional network variables is that the access to any network ceases to be important, both in substantive and statistical terms, in most cases—whereas some of the more disaggregated (or more specific) network access variables "take over" their importance in a few cases. For our main health knowledge variable (diarrhea treatment in children), for example, it turns out that having access to a doctor in one's social network is the most important, being associated with about a 15 percentage-point increase in that specific health knowledge (though only marginally statically significant, at a level of statistical significance of 10 percent or less).

We also perform a sensitivity analysis, by applying propensity score matching as an alternative estimation method (Table 7). We do this by using the core set of explanatory variables in the matching (that is, all explanatory variables except social network access), and then using any health network access and any education network access as our treatment variables, respectively. From the results, this method seems to bring out more of the social network effect. For our main health knowledge variable of interest (diarrhea treatment), for example, knowing any health person or knowing any education person is associated with a 2 and a 6.4 percentage-points increase in knowing about the correct way to treat diarrhea in children regarding their water intake. In turn, this illustrates that not only access to health networks are important for health knowledge, access to education networks are also important—and sometimes, like in this case, even more so.

Table 4. OLS/LPM and 2SLS/IV Social Network Coefficients: Only Any Health Network (With All Core Variables Included)

	Milk 1	Sterilization	Milk 2	Smoke	Diarrhea	Menstruation	Score Index
OLS/LPM:							
Any health	0.019+	0.058***	0.015	0.011	0.054***	-0.018*	0.140***
	[0.015]	[0.015]	[0.013]	[0.012]	[0.015]	[0.010]	[0.043]
R^2	0.251	0.226	0.222	0.304	0.29	0.229	0.312
IV/2SLS:							
Any health	0.061+	0.083*	0.003	0.070++	0.202***	-0.062**	0.357**
	[0.044]	[0.048]	[0.041]	[0.043]	[0.046]	[0.029]	[0.155]
N	16,468	16,468	16,468	16,468	16,468	16,468	16,468

Notes: Robust Huber-White (Huber, 1967; White, 1980) standard errors, adjusted for within-cluster correlation/clustering (Wooldridge, 2010), in brackets under parameter estimates. ***: statistically significant at 1 percent; **: statistically significant at 5 percent; *: statistically significant at 10 percent; ++: statistically significant at 15 percent +: statistically significant at 20 percent. All specifications include district fixed effects. Additional controls include a dummy for high caste, age cohort dummies, and dummies for educational attainment, information exposure, and access to health facilities in the community.
Source: 2004/05 India Human Development Survey (IHDS).

Table 5. OLS/LPM and 2SLS/IV Social Network Coefficients: Adding Any Education Network (With All Core Variables Included)

	Milk 1	Sterilization	Milk 2	Smoke	Diarrhea	Menstruation	Score Index
OLS/LPM:							
Any health	0.003	0.057***	0.015	0.004	0.031**	-0.015+	0.096**
	[0.016]	[0.015]	[0.014]	[0.012]	[0.016]	[0.011]	[0.041]
Any edu	0.044***	0.001	-0.001	0.018*	0.061***	-0.008	0.116***
	[0.014]	[0.013]	[0.012]	[0.010]	[0.015]	[0.010]	[0.033]
R^2	0.253	0.226	0.222	0.305	0.292	0.229	0.313
N	16,468	16,468	16,468	16,468	16,468	16,468	16,468
IV/2SLS:							
Any health	0.081*	0.06	-0.014	0.078++	0.178***	-0.029	0.353**
	[0.048]	[0.049]	[0.048]	[0.048]	[0.049]	[0.037]	[0.148]
Any edu	-0.033	0.039	0.029	-0.014	0.041	-0.056++	0.006
	[0.047]	[0.047]	[0.049]	[0.047]	[0.044]	[0.036]	[0.134]
N	16,468	16,468	16,468	16,468	16,468	16,468	16,468

Notes: Robust Huber-White (Huber, 1967; White, 1980) standard errors, adjusted for within-cluster correlation/clustering (Wooldridge, 2010), in brackets under parameter estimates. ***: statistically significant at 1 percent; **: statistically significant at 5 percent; *: statistically significant at 10 percent; ++: statistically significant at 15 percent +: statistically significant at 20 percent. All specifications include district fixed effects. Additional controls include a dummy for high caste, age cohort dummies, and dummies for educational attainment, information exposure, and access to health facilities in the community.
Source: 2004/05 India Human Development Survey (IHDS).

Table 6. OLS/LPM and 2SLS/IV Social Network Coefficients: Full Specification With All Network Variables Included (With All Core Variables Included)

	Milk 1	Sterilization	Milk 2	Smoke	Diarrhea	Menstruation	Score Index
OLS/LPM:							
Any health	0.024	0.047**	0.026	0.024+	0.029+	-0.052***	0.098*
	[0.020]	[0.021]	[0.022]	[0.018]	[0.022]	[0.016]	[0.056]
Any edu	0.061*	0.006	-0.035	0.039*	0.083**	-0.028+	0.125*
	[0.033]	[0.030]	[0.030]	[0.020]	[0.033]	[0.020]	[0.070]
Any doctor	-0.015	0.027+	-0.007	-0.019	-0.014	0.054***	0.026
	[0.023]	[0.020]	[0.021]	[0.018]	[0.024]	[0.016]	[0.053]
Any Teach/Princ	-0.013	-0.012	0.018	-0.016	-0.022	0.006	-0.039
	[0.030]	[0.028]	[0.028]	[0.018]	[0.029]	[0.017]	[0.063]
Health, same jati	-0.031*	-0.026	-0.011	-0.018	0.039*	-0.005	-0.052
	[0.019]	[0.023]	[0.020]	[0.017]	[0.022]	[0.015]	[0.054]
Edu, same jati	-0.011	0.01	0.038**	-0.013	-0.004	0.027**	0.047
	[0.017]	[0.018]	[0.016]	[0.012]	[0.020]	[0.013]	[0.043]
R^2	0.253	0.226	0.223	0.305	0.292	0.231	0.313
IV/2SLS:							
Any health	0.019	0.061	-0.057	0.03	0.057	-0.005	0.104
	[0.079]	[0.065]	[0.086]	[0.072]	[0.083]	[0.063]	[0.197]
Any edu	0.009	-0.043	-0.075	-0.076	0.052	-0.233***	-0.365+
	[0.111]	[0.085]	[0.123]	[0.080]	[0.100]	[0.078]	[0.271]
Any doctor	0.119++	0.039	0.104+	0.047	0.151*	-0.044	0.416**
	[0.074]	[0.065]	[0.079]	[0.064]	[0.086]	[0.061]	[0.210]
Any Teach/Princ	-0.005	0.114++	0.077	0.111++	0.012	0.170**	0.479*
	[0.100]	[0.078]	[0.115]	[0.074]	[0.097]	[0.073]	[0.265]
Health, same jati	-0.117+	-0.133*	-0.109+	0	0.011	0.02	-0.329++
	[0.085]	[0.070]	[0.084]	[0.078]	[0.084]	[0.055]	[0.224]
Edu, same jati	-0.119*	-0.052	0.073	-0.085++	-0.069	0.088*	-0.164
	[0.065]	[0.066]	[0.074]	[0.056]	[0.068]	[0.050]	[0.173]
N	16,468	16,468	16,468	16,468	16,468	16,468	16,468

Notes: Robust Huber-White (Huber, 1967; White, 1980) standard errors, adjusted for within-cluster correlation/clustering (Wooldridge, 2010), in brackets under parameter estimates. ***: statistically significant at 1 percent; **: statistically significant at 5 percent; *: statistically significant at 10 percent; ++: statistically significant at 15 percent +: statistically significant at 20 percent. All specifications include district fixed effects. Additional controls include dummy for high caste, age cohort dummies, and dummies for educational attainment, information exposure, and access to health facilities in the community.
Source: 2004/05 India Human Development Survey (IHDS).

Table 7. Sensitivity Analysis: Propensity Score Matching Results For Any Health and Any Education Networks Treatment Impacts (Nearest Neighbor)

	Knows Any Health Person	*Knows Any Education Person*
Milk drinking during pregnancy	0.027** [0.017]	0.025* [0.013]
Physical weakness of men after sterilization	0.043*** [0.015]	-0.016++ [0.015]
Goodness of the first (thin) milk for the baby	0.002 [0.014]	0.032*** [0.015]
Goodness of smoke from wood/dung burning	0.039*** [0.012]	0.059*** [0.012]
Treatment of diarrhea in children, re water intake	0.020** [0.020]	0.064*** [0.016]
Menstruation, re "safe period"	-0.003 [0.014]	0.005+ [0.011]
Combined (score-) index	0.126*** [0.037]	0.168*** [0.041]
N	16,468	16,468

Notes: Explanatory/matching variables include all the core explanatory variables (i.e., all explanatory variables except network variables). ***: statistically significant at 1 percent; **: statistically significant at 5 percent; *: statistically significant at 10 percent; ++: statistically significant at 15 percent +: statistically significant at 20 percent.
Source: 2004/05 India Human Development Survey (IHDS).

Question 5: *What explains the observed caste gap in health knowledge—and is there evidence that this may be due to discrimination against the low caste?*

(i) Overall health knowledge gap decompositions:

A couple of results stand out particularly strongly from the results of the three-fold decompositions (Table 8, top panel). First, the endowments increase the caste health knowledge gap overall in all cases (and that both statistically and substantively significantly so), indicating that high caste women have relatively more favorable observable characteristics—that is, they have more (and possibly also better) education, are more exposed to information relevant for health knowledge production, and have more access to social networks (this will be examined more closely when considering the detailed

decompositions in the next sub-section). Second, while the returns to these characteristics decrease the gaps in most cases (though not always statistically significantly so), indicating that low caste women have higher returns to characteristics overall, this is not the case for our main health knowledge measure (diarrhea in children): here, the returns to characteristics work to increase the health knowledge gap.

Moving to the two-fold decompositions, high caste women on average have better health knowledge related characteristics (such as educational attainment, information exposure, and social network access) as indicated by the positive sign in the explained part—which in turn serves to increase the caste health knowledge gap—whereas the unexplained part (capturing all the factors that cannot be attributed to differences in observed characteristics) mostly accounts for a somewhat smaller share of the caste health knowledge differential (Table 8, bottom panel). Again the unexplained part mostly works to decrease the gap—except for our preferred health knowledge measure, where the unexplained part explains almost all the gap.

Notably—as can be seen from the results from the sensitivity analysis shown in Appendices C and D—these results are quite robust to whether the decomposition is performed from low caste women's viewpoint (i.e., using high caste endowments and returns) or whether the decomposition is performed from high caste women's viewpoint (i.e., using low caste endowments and returns) for the three-fold decompositions or from any of the many different possibilities of specifying the "absence of discrimination" group in the two-fold decompositions.

(ii) Detailed health knowledge gap decompositions:

Examining the detailed caste health knowledge decompositions allows us to assess in more detail what the individual components of the overall health knowledge gaps are, in terms of specific (groups of) explanatory variables. From Tables C.XX and C.YY (Appendix C) the main component of the endowment (three-fold) and explained (two-fold) parts of the overall gaps is education. For our preferred health knowledge measure we note how knowing a doctor is very important, as well, explaining about half of the observed gap.

Table 8. Overall Health Knowledge Gap Decompositions: Three- and Two-fold

	(1) Milk drinking during pregnancy	(2) Physical weakness of men after sterilization	(3) Goodness of the first (thin) milk for the baby	(4) Goodness of smoke from wood/ dung burning	(5) Treatment of diarrhea in children, re water intake	(6) Menstruation, re "safe period"	(7) Combined (score-) index
Three-fold:							
Endowments	0.029**	0.040***	0.047***	0.041***	0.102***	0.037***	0.295***
	[0.014]	[0.012]	[0.013]	[0.011]	[0.014]	[0.010]	[0.036]
Coefficients	-0.026	-0.051**	-0.146***	-0.007	0.119***	-0.048**	-0.159**
	[0.026]	[0.024]	[0.023]	[0.021]	[0.025]	[0.019]	[0.066]
Interaction	0.055**	0.084***	0.158***	-0.016	-0.107***	0.037*	0.211***
	[0.028]	[0.025]	[0.024]	[0.023]	[0.027]	[0.020]	[0.071]
Two-fold:							
Explained	0.083***	0.124***	0.205***	0.025	-0.005	0.073***	0.506***
	[0.025]	[0.023]	[0.022]	[0.021]	[0.024]	[0.019]	[0.065]
Unexplained	-0.026	-0.051**	-0.146***	-0.007	0.119***	-0.048**	-0.159**
	[0.026]	[0.024]	[0.023]	[0.021]	[0.025]	[0.019]	[0.066]
N	16,468	16,468	16,468	16,468	16,468	16,468	16,468

Notes: Decompositions are from low caste females' viewpoint—i.e., using high caste endowments and returns (sensitivity analysis using reverse decompositions for the three-fold decomposition and several alternative weights given to high caste relative to low caste used in determining the reference coefficients for the two-fold decompositions are reported in Appendices C and D, respectively). Values in brackets are robust Huber-White (Huber, 1967; White, 1980) standard errors. ***: statistically significant at 1 percent; **: statistically significant at 5 percent; *: statistically significant at 10 percent; ++: statistically significant at 15 percent +: statistically significant at 20 percent.
Source: 2004/05 India Human Development Survey (IHDS).

6. Conclusion

This paper examines the caste health knowledge gap in India in terms of its prevalence, magnitude and determinants using a recent data set and thereby add to the emerging literature on caste and health knowledge.

Estimation of raw caste health knowledge gaps and overall and detailed earnings decompositions leads to four main results: (1) education, information exposure, and social network access are all strongly associated with increased health knowledge; (2) the presence of a substantively large health knowledge caste gap (favoring high caste women); (3) evidence that the endowments and (though only for our preferred health knowledge measure, on the correct treatment of diarrhea in children) the returns to characteristics increase the health knowledge gaps—indicating that high caste women have higher education, have greater information exposure, and have better access to social networks; (4) while observed individual characteristics explain part of the gaps, a substantial part of the health knowledge gap is left unexplained.

These results have strong policy implications, consistent as they are with the presence of discrimination towards low caste women in the context of health knowledge. In particular, the continued presence of a caste health knowledge gap—especially regarding the correct treatment of diarrhea in children regarding their water intake—is likely to lead to continued child mortality for children from low caste backgrounds. Notably, these are deaths that could have been averted, had the mothers only been taught the arguably simple—and at the same time also relatively cheap—measure of increasing the water intake for their children when experiencing diarrhea. In turn, this points towards the importance of continued attention towards education, institutions and economic policy for decreasing the caste gap in India—notably through increased attention towards the education system and public provision of health information campaigns.

References:

Borooah, Vani and Sriya Iyer (2005) "Vidya, Veda, and Varna: The influence of religion and caste on education in rural India," *The Journal of Development Studies* 41(8): 1369-1404.

Becker, Gary S. (1964) *Human Capital*, Chicago: University of Chicago Press.

Blinder, A.S. (1973) "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources* 8: 436-455.

Card, David (2001) "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69(5): 1127-1160.

Chin, Aimee and Nishith Prakash (2011) "The redistributive effects of political reservation for minorities: Evidence from India," *Journal of Development Economics* 96 (2): 265-277.

Cotton, J. (1988) "On the Decomposition of Wage Differentials," *Review of Economics and Statistics* 70: 236-243.

Desai, Sonalde, Reeve Vanneman, and National Council of Applied Economic Research, New Delhi (2010). India Human Development Survey (IHDS), 2005 [Computer file]. ICPSR22626-v8. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-06-29. doi:10.3886/ICPSR22626.v8

Heckman, James J., Lance J. Lochner, and Petra E. Todd (2008) "Earnings Functions and Rates of Return," *Journal of Human Capital* 2(1):1-31.

Grossman, Michael (1972) "On the Concept of Health Capital and the Demand for Health," *Journal of Political Economy* 80(2): 223-55.

Huber, P. J. (1967) "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in: Lucien M. Le Cam and Jerzy Neyman (Eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability Vol. 1*, Berkeley, CA: University of California Press.

Jann, Ben (2008) "The Blinder-Oaxaca decomposition for linear regression models," *The Stata Journal* 8(4): 453-479.

Kenkel, Donald S (1991) "Health Behavior, Health Knowledge, and Schooling," *Journal of Political Economy* 99(2): 287-305.

Mincer, Jacob (1974) *Schooling, Experience and Earnings*, New York: National Bureau of Economic Research.

Neumark, D. (1988) "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination," *Journal of Human Resources* 23: 279-295.

Oaxaca, R. (1973) "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review* 14: 693-709.

Oaxaca, R.L., Ransom, M.R. (1998) "Calculation of approximate variances for wage Decomposition differentials," *Journal of Economic and Social Measurement* 24: 55-61.

Oaxaca, R.L. and M.R. Ransom (1999) "Identification in Detailed Wage Decompositions," *Review of Economics and Statistics* 81: 154-157.

Reimers, C.W. (1983) "Labor Market Discrimination against Hispanic and Black Men," *Review of Economics and Statistics* 65: 570-579.

White, H. (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48(4):817–830.

Wooldridge, Jeffrey M. (2010) *Econometric Analysis of Cross-Section and Panel Data*, Second Edition, Cambridge, Massachusetts: The MIT Press.

Yun, Myeong-Su (2005) "A Simple Solution to the Identification Problem in Detailed Wage Decompositions," *Economic Inquiry* 43: 766-772.

APPENDIX A: Descriptive Statistics

Table A1. Descriptive Statistics for Estimation Sample

	High caste women:		Low caste women:	
	Mean	Std Dev	Mean	Std Dev
Dependent variables:				
Milk drinking during pregnancy	0.769	0.421	0.712	0.453
Physical weakness of men after sterilization	0.334	0.472	0.261	0.439
Goodness of the first (thin) milk for the baby	0.756	0.429	0.697	0.460
Goodness of smoke from wood/dung burning	0.798	0.401	0.781	0.414
Treatment of diarrhea in children, re water intake	0.604	0.489	0.490	0.500
Menstruation, re "safe period"	0.154	0.361	0.128	0.334
Combined (score-) health knowledge index	3.415	1.239	3.068	1.258
Explanatory variables:				
Age cohorts:				
15-19	0.025	0.156	0.042	0.200
20-24	0.137	0.344	0.159	0.366
25-29	0.189	0.391	0.181	0.385
30-34	0.199	0.399	0.191	0.393
35-39	0.197	0.398	0.196	0.397
40-44	0.149	0.357	0.137	0.344
45-49	0.104	0.305	0.095	0.293
Educational attainment:				
No education	0.354	0.478	0.620	0.486
Some education (pri incomplete)	0.077	0.266	0.081	0.273
Primary	0.227	0.419	0.144	0.351
Middle/some secondary	0.270	0.444	0.130	0.336
Higher secondary	0.046	0.208	0.017	0.129
Tertiary	0.026	0.161	0.008	0.090
Information exposure:				
Reads newspapers regularly	0.096	0.294	0.032	0.175
Watches tv regularly	0.471	0.499	0.261	0.439
Network access:				
Knows any health person	0.407	0.491	0.289	0.453
Knows any education person	0.527	0.499	0.362	0.481
Knows any doctor	0.332	0.471	0.228	0.419
Knows any teacher/principal	0.469	0.499	0.326	0.469
Health person is of same jati	0.173	0.378	0.075	0.263
Education person is of same jati	0.316	0.465	0.147	0.354
Access to health facilities in village:				
Health Sub-center in community	0.420	0.494	0.446	0.497

Primary Health Center in community	0.141	0.348	0.170	0.375
Community Health Center	0.032	0.177	0.022	0.146
Government Maternity Center	0.029	0.167	0.057	0.232
Govt. Communicable Disease Facility (e.g., TB)	0.028	0.164	0.052	0.223
N	3,707		12,761	

Notes: Calculations incorporate sampling weights and clustering (Wooldridge, 2010).
Source: 2004/05 India Human Development Survey (IHDS).

APPENDIX B: Specification Tests

[TO BE COMPLETED!]

Table B1. Specification Tests for 2SLS/IV Health Knowledge Regressions: Predictive Power of Identifying Instruments (First Stage); Endogeneity (Second Stage)

	(1) Core set of explanatory variables	*(2) Adding "Any health network"*	*(3) Additionally adding "Any education network"*	*(4) Additionally adding all remaining network variables*
Joint F-test of predictive power of IVs:				
Reads newspapers regularly				
Watches tv regularly				
Any health network				
Any education network				
Any doctor				
Any teacher/principal				
Any health: same jati				
Any education: same jati				
Wu (1973)-Hausman (1978) endogeneity test:				
Milk drinking during pregnancy				
Physical weakness of men after sterilization				
Goodness of the first (thin) milk for the baby				
Goodness of smoke from wood/dung burning				
Treatment of diarrhea in children, re water intake				
Menstruation, re "safe period"				
Combined (score-) health knowledge index				
N				

Notes: Terms in brackets are the p-values of the corresponding test-statistic. The tests employ robust Huber-White (Huber, 1967; White, 1980) standard errors and also adjust for within-community correlation/clustering (Wooldridge, 2010). All estimations include cluster fixed-effects and remaining explanatory variables similar to those used for the main estimations.
Source: 2004/05 India Human Development Survey (IHDS).