Online Appendix for *Jobs for Sale: Corruption and Misallocation in Hiring* (Weaver)

A Online Supplemental Appendix Tables and Figures

Table A.1: Hiring as a supervisor and wealth

	Applicants		Full Sample	
	(1)	(2)	(3)	(4)
Wealth	0.239	0.322	0.130	0.141
	(0.0663)	(0.0820)	(0.0257)	(0.0272)
Intercept	0.0597		0.00193	
	(0.0382)		(0.0128)	
Cluster FEs	No	Yes	No	Yes
Observations	339	338	986	986

Notes: This table reports the relationship between wealth and whether an individual was hired as a supervisor. The dependent variable is a binary variable denoting whether the individual was hired as supervisor. Columns (2) and (4) include cluster fixed effects to account for unobserved heterogeneity. Since there is one cluster with only one applicant, this results in one observation being dropped in column (2) relative to column (1). Wealth is measured as the individual's percentile rank in the wealth distribution among all respondents (0 is poorest, 1 is wealthiest).

Table A.2: Supervisor performance index and health services delivery

	Institutional Delivery	Newborn Check-ups	Nutritional Counseling	DOTS provider
Predicted SPI	.0078	.0053	.0059	.0018
	(.0034)	(.0017)	(.0015)	(.0011)
Observations	917	917	917	917
Initial Mean	2.598	0.784	0.109	0.200
Effect Size (1 SD)	0.155	0.106	0.119	0.0353

Notes: This table studies how performance changes over a 20 month time period under a supervisor with a higher value of the characteristics in the SPI index. Coefficients are equal to the average monthly improvement on the outcome of interest for a CHW whose supervisor has a one SD higher value of SPI. Initial Mean is equal to the mean value for delivery of that service across CHWs for the first three months of data. Effect Size (1 SD) translates the coefficient into the estimated relationship with service delivery over the 20 months of data. Column (1) refers to the number of institutional deliveries that the community health worker assisted in a month. Column (2) is the fraction of newborn children upon whom the community health worker conducted a check-up in the month. Column (3) is the percentage of pregnant women in their catchment area that the community health worker visited and provided nutritional counseling in the month. Column (4) is a binary variable for whether the community health worker was serving as a tuberculosis treatment provider during this month. Since SPI is a constructed regressor, clustered standard errors will overestimate the true level of precision. To incorporate the additional uncertainty from construction of the index, I use clustered bootstrapped standard errors, where SPI is reconstructed in each bootstrap sample based on that sample of data.

Table A.3: Comparison of Measures of Supervisor Quality

	Functionality Score	Functionality Score	Functionality Score	Functionality Score	Functionality Score
Supervisor SPI	0.54				0.51
	(0.059)				(0.056)
Performance Evaluation		0.10			0.044
		(0.039)			(0.034)
CHW Rating			0.74		0.78
			(0.387)		(0.328)
Process Rating				0.11	-0.10
				(0.080)	(0.078)
Observations	917	917	917	917	917
R^2	0.162	0.039	0.028	0.016	0.185

Notes: This table examines the extent to which SPI captures supervisor quality, as there may be unobserved elements of supervisor quality that SPI does not measure. Column (1) regresses average monthly change in the functionality score of CHWs after the supervisor is hired on SPI, while columns (2), (3) and (4) regress the same outcome on three alternative evaluations of the quality of the supervisor. In order, these evaluations are: performance evaluations of the the supervisors by individuals overseeing the program (column 2), ratings of the supervisor by the CHWs they supervise (column 3), and an index of process measures of supervisor performance (frequency of interactions with CHWs, supervisory tasks completed over the past two months, column 4). Columns (5) tests whether the alternative measure contain additional information predictive of service delivery changes after SPI is accounted for.

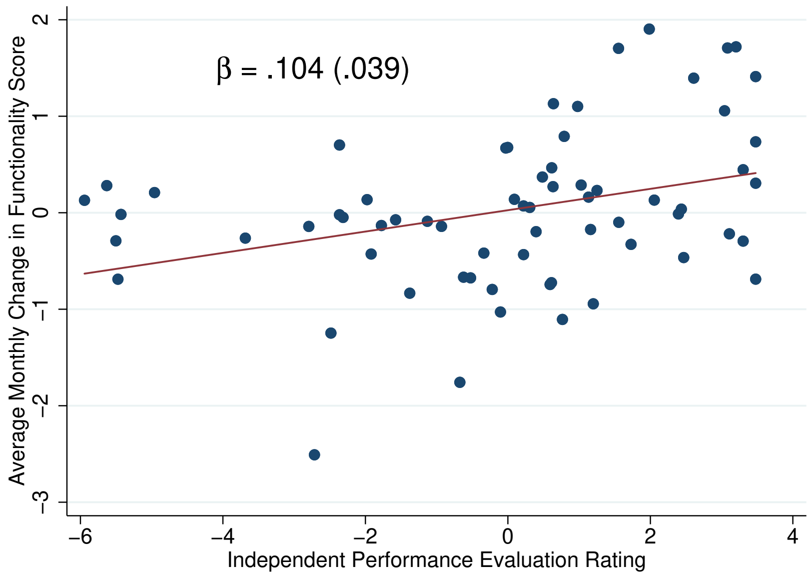
Table A.4: Placebo test for cluster-level wealth-quality correlations

	Coefficient	Standard Error	p-value
Wealth	.00011	(.028)	[1]
Supervisor Performance Index	.042	(.36)	[.91]
Raven's Score	.0082	(.26)	[.97]
Education	.26	(.35)	[.47]
Reading Skill	.029	(.23)	[.9]
Writing Skill	-.016	(.26)	[.95]
Health Knowledge	-.0012	(.043)	[.98]
Government Hiring System Points	-.081	(.29)	[.78]
Monthly Deliveries (Baseline)	-.26	(.26)	[.32]
Initial Newborn Visits (Admin)	-.049	(.078)	[.53]
Initial Nutrition Counseling (Admin)	-.066	(.064)	[.31]
Initial Deliveries (Admin)	-.29	(.26)	[.27]
Initial DOTS Provider (Admin)	-.079	(.059)	[.18]
Initial Functionality Score (Admin)	-8.1	(5.4)	[.14]
Monthly Deliveries (Pre-trend)	-.039	(.11)	[.73]
Number of CHWs in cluster	1.9	(1.2)	[.11]
Number of applicants	1.1	(.63)	[.085]
Village population	-68	(43)	[.12]
Observations	986		
Joint p-value		.65	

Each row gives the coefficient, standard error, and p-value from a regression of individual CHW characteristics on the correlation of wealth and quality within the cluster. "Government Hiring System Points" is the number of points that the CHW had under the intended government counterfactual hiring system. "Monthly Deliveries (Baseline)" comes from the baseline survey done with CHWs prior to the hiring of the supervisors, while the other measures of service delivery are the value of that outcome in first time that the CHW is observed in the government administrative data. "Monthly Deliveries (Pre-trend)" is the trend in deliveries for that CHW over the six months prior to the start of supervisors. Village population is the population of the village that the CHW works in. Standard errors are in parentheses and clustered at the supervisor level, while p-values are in brackets. Joint p-value comes from an F-test of joint significance.

Figure A.1: Timeline of Project and Data Collection

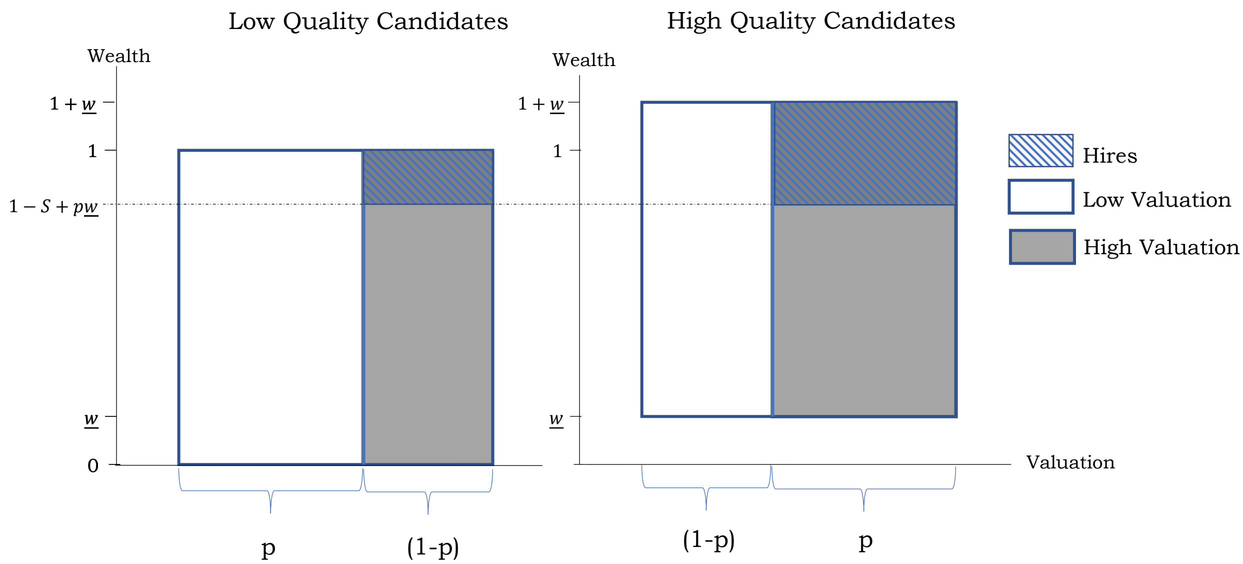
Figure A.2: Performance Evaluations and Administrative Data



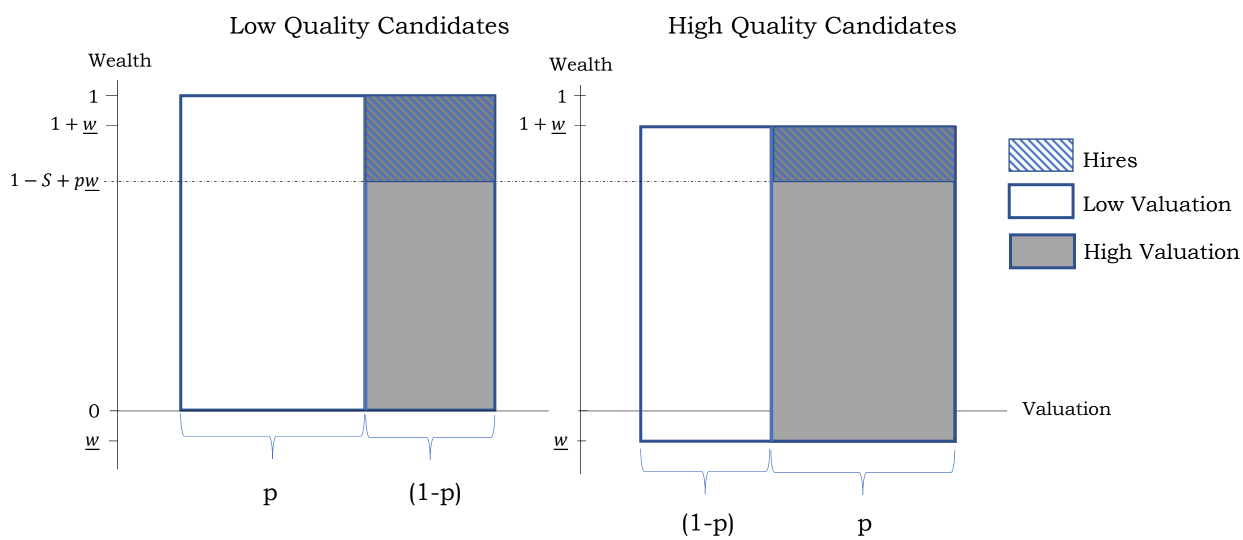
Notes: This figure compares administrative data on the performance of CHWs supervised by a given supervisor against independent performance evaluations of those supervisors to check for manipulation of the administrative data. On the y-axis is the average monthly change in functionality score for their CHWs after the supervisor was hired over the 20 months of administrative data. The independent performance evaluations are on the x-axis and come from individuals overseeing the supervisors who were not involved in hiring. The performance evaluations were along nine dimensions: overall quality of supervisor, quality of meetings with CHWs, health knowledge, overall intelligence, selflessness, interest in improving health outcomes, competitiveness, and desire to help CHWs improve. Independent Performance Evaluation Rating is the first principal component of those evaluations.

Figure A.3: Model illustration

(a) Positive correlation between wealth and quality, positive correlation between valuation and quality

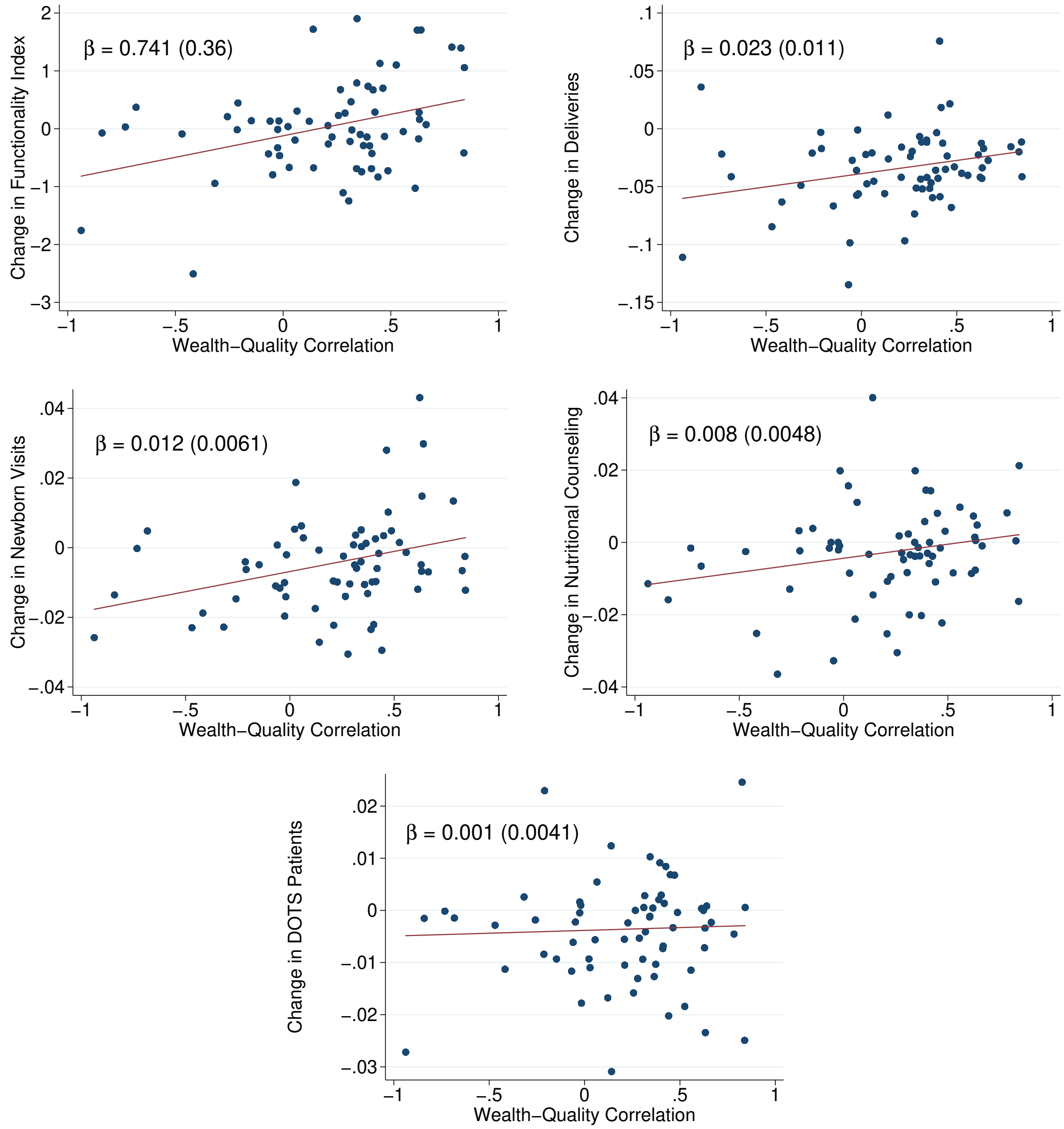


(b) Negative correlation between wealth and quality, positive correlation between valuation and quality



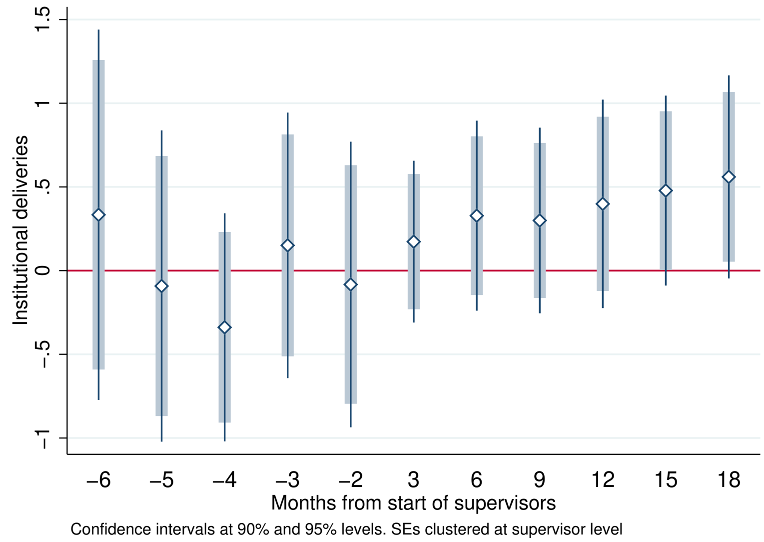
Notes: This figure illustrates the parameters of the model in section IV. In the model, there is measure 1 of both low and high quality candidates applying for measure S jobs. Candidate wealth is uniformly distributed between $[0, 1]$ for low quality candidates and $[\underline{w}, 1+\underline{w}]$ for high quality candidates, where $-1 < \underline{w} < 1$. Candidates either have a high or low valuation for the job, where a proportion p of high quality candidates and $(1 - p)$ of low quality candidates place a high valuation on the job. The left square represents the low quality candidates and the right square is the high quality candidates. The x-axis represents valuation (which I discretize into low and high for simplicity in the model), the y-axis represents wealth, and candidate valuation/wealth is uniformly distributed over the square. The low valuation candidates are unshaded, while the high valuation candidates are shaded. Those candidates who are hired are marked by the diagonal lines, i.e. those above a wealth of $1 - S + p\underline{w}$.
In the top figure, there is a positive correlation between wealth and quality ($\underline{w} > 0$) and a positive correlation between valuation and quality ($p > 0.5$), while in the bottom, there is a negative correlation between wealth and quality ($\underline{w} < 0$) and a positive correlation between valuation and quality ($p > 0.5$). As the correlation betweeen wealth and quality decreases, the proportion of hires who are high quality also decreases.

Figure A.4: Wealth-Quality Correlations and Service Delivery Outcomes



Notes: This figure examines the relationship between service delivery outcomes and the correlation of SPI and wealth within a given cluster. Each panel plots the average monthly change in a measure of service delivery in the cluster after the hiring of the supervisor against the correlation of wealth and SPI within that cluster. The first panel is of the "functionality score", which aggregates the other service delivery outcomes into a rating between 0 and 100 via a government formula.

Figure A.5: Event study for the cluster-level correlation of wealth and SPI



Notes: This figure plots the event study for how institutional deliveries are related to the correlation of wealth and SPI of the hired supervisor. It plots the coefficients from a regression of the number of institutional deliveries assisted by a CHW on the interactions of whether the data is from a particular month and the correlation of wealth and SPI of the hired supervisor. Note that it is really a combination of two event studies from different data sets: for the months prior to the start of supervisors, I am using data from surveys of CHWs, while for the months following the start of supervisors, I use the administrative data on deliveries. I combine the months following the start of supervisors into three month bins to reduce the amount of noise in the estimates. Confidence intervals are at the 90% and 95% levels. Standard errors are clustered at the supervisor cluster level.

B Supplemental Material on Data Collection

B.1 Survey Details

For both rounds of surveying, CHWs and supervisors were contacted via phone and made appointments to take the survey at a convenient central location. Respondents were paid 150% of average CHW daily earnings. They also had the potential to earn up to an additional 75% based on their performance in behavioral games, but they were not told this prior to arrival. Since the survey took between 1-2 hours, and payment exceeded the typical daily wage rate, refusal rates were very low. All surveyors were female since the CHWs are female, and most had previous experience surveying this population of CHWs.

The first round of surveying occurred after supervisors had been hired, but a month before they had started their new duties. The first survey was focused on the work of CHWs over the preceding six months, as well as administering a test of health knowledge and numerous psychometric instruments. It also asked about the hiring of supervisors. The second round of surveying was six months after supervisors began their work and attempted to interview all CHWs and supervisors. This survey focused on the performance of supervisors and CHWs, as well as administering two tests of general ability (Raven's Progressive Matrices; digit span memory test), a test of health knowledge, a behavioral game measuring pro-social preferences, and a behavioral game measuring honesty. We were able to contact 96.4% of the sample frame, and of those contacted, 92% were administered the survey.[30] The remainder of the section describes tests and behavioral games used during data collection.

Ability (Problem solving): The Raven's Progressive Matrices measure general cognitive ability and have been used in hundreds of academic papers, as well as by some government agencies (e.g. Dal Bó et al. (2013); similar problems have been included in Mexican and Indian civil service exams). The test consists of a series of visual patterns of abstract shapes. From each pattern, a piece is missing, and respondents must identify the missing piece from a list of options. To induce effort, respondents were paid for each correctly answered problem and could earn up to a third of the prevailing daily wage. The set of 12 matrices were taken from the Advanced Progressive Matrices, Set I, as published by Pearson Clinical.

Short-term Memory: In the digit span memory test, surveyors recite a string of digits (e.g. 1-8-3-4-5) to the respondents and ask them to repeat it in same order. Following this, they are given a second string containing the same number of digits, and again have a single opportunity to give a correct answer. If either response is correct, then the number of digits increases by one, and the process repeats until the respondent cannot successfully repeat either opportunity for a given number of digits. The longest number of digits correctly repeated is their score.

Pro-social Preferences: Pro-social preferences were measured via a modified dictator game, which other studies have found is predictive of real-world pro-social actions (e.g. Lagarde and Blaauw, 2014). Respondents were informed that after the survey was completed, we would select

[30] Among those that we were not able to contact, some have likely discontinued their work as a CHW. However, in cases where we were unable to confirm this, I leave them in the sample frame. In cases where it was not possible to contact a particular CHW by phone, we attempted to contact them via other health workers who lived near them. Outright refusal rates were very low (0.5%), with most attrition due to being out of town or family obligations.

sixteen respondents and give them an an amount approximately equal to a third of their average monthly earnings. If they desired, they could donate some fraction of this to a local orphanage, but they had to decide this at the time of the survey, prior to finding out if they had won. Respondents were given this amount in fake local notes to split between envelopes marked "donation" and "self".

Dishonesty: Honesty was measured using a modified version of a behavioral game from Hanna and Wang (2017). After completing the survey and other games, CHWs were given a dice and told to roll it 40 times, noting their rolls on a sheet. They were told that for each roll of 5 or 6, they would receive a payment, but for any other roll, they would not receive anything. They could earn slightly less than half the prevailing daily wage if they reported all 5's or 6's. Respondents have an incentive to act dishonestly by reporting a larger number of 5's and 6's, and can do so without fear of repercussion, since it is never possible for others to know conclusively whether they cheated. This game was played after the survey to avoid priming on dishonesty. 58% reported winning rolls above the 95^{th} percentile of what would be expected by chance.

Table B.1: Variables included in LASSO for creation of SPI

Category (# unique variables)	**Description**
Health Knowledge (1)	Score on a 30 question test of health knowledge
Digit span Test (1)	Score on digit span memory test
Raven's Matrices (1)	Score on Raven's Progressive Matrices
Education (1)	Years of Education
Honesty Game (1)	Number of high dice rolls in honesty game
Management Experience (1)	Number of employees previously managed
Reading ability (1)	Score on reading test during survey
Writing ability (1)	Score on writing test during survey
Pro-Sociality/ Generosity (1)	Charitable donation in dictator game
Public Service Motivation (1)	Psychometric index measuring Public Service Motivation
Motivation (2)	Extrinsic/Intrinsic motivation scales
Big Five Personality Index (5)	Big Five Personality Index
Time Worked (1)	Hours worked per week
CHW Clients (1)	Total number of clients (as a CHW)
Peer Rating (3)	Rating of performance as a CHW, health knowledge, motivation
CHW Tasks (1)	Number of tasks carried out as a CHW
Medical Advice (1)	Frequency of giving advice on ailments (as a CHW)
DOTS provider (1)	Number of tuberculosis patients serving (as a CHW)
Institutional Deliveries (1)	Women brought to give birth at hospital (as a CHW)
CHW Performance (1)	Summary measure of performance (as a CHW)
Job Satisfaction as CHW (1)	5-point scale measuring their job satisfaction (as a CHW)

B.2 Validations of the Administrative Data on Health Services Delivery

One concern with the administrative data on health services delivery is that the supervisors might manipulate it. Note that there is no direct incentive for them to do so since this data does not affect their financial compensation: they are paid a fixed monthly amount independent of CHW performance. Nonetheless, I run three checks for manipulation.

First, I collected a second set of measures of supervisor performance from individuals overseeing the program but who were not part of hiring process. Supervisors could not manipulate this data since they did not know it was being collected. These individuals ranked the supervisors on 9 performance measures (e.g. overall performance, motivation) based on field visits and interactions with CHWs. They have no incentive to misreport the performance of one supervisor relative to another; they knew this was only for research purposes and would have no impact on the supervisor. They also do not have supervisor-level administrative data on health services delivery, so their ratings are not mechanically related to the administrative data. Figure A.2 and column (1) of table B.2 compare a PCA index of the 9 collected performance measures against the administrative data (average monthly change in functionality score for CHWs under the supervisor). The strong relationship ($p < 0.01$) indicate that the administrative data reflect ground reality rather than manipulation.

Second, I check how the administrative data compares to independent survey data. During the second survey round, CHWs were asked about how many deliveries they had assisted in each month for the last six months. Column (3) regresses the number of deliveries in the administrative data in a given month on the corresponding survey data. The recall periods differ slightly, so it is not surprising that the two do not align completely. However, the relationship is strong ($p < 0.001$), arguing against major manipulation. Third, if manipulation explained my results, then we would expect the relationship between the administrative and independent data to be weaker for high SPI supervisors (negative interaction term). Columns (2) and (4) shows that the relationship between the administrative and independent measures is not intermediated by the SPI of their supervisor.

Table B.2: Validation of Administrative Data

	Functionality Score	Functionality Score	Deliveries (Admin)	Deliveries (Admin)
Evaluation Rating	0.104	0.103		
	(0.0387)	(0.0379)		
Evaluation Rating X Supervisor SPI		-0.0114		
		(0.0814)		
Deliveries (Survey)			0.375	0.374
			(0.0259)	(0.0260)
Deliveries (Survey) X Supervisor SPI				0.0182
				(0.0242)
Dependent mean	.059	.059	2.1	2.1
Observations	917	917	3,005	3,005

Notes: This table cross-checks the administrative data. Column (1) regresses adminstrative data for the CHWs supervised by a given supervisor against independent performance evaluations of those supervisors to check for manipulation. Column (3) reports estimates of the relationship between data on deliveries in a given month in administrative and survey data. Columns (2) and (4) test whether supervisor SPI intermediates either relationship, which could be consistent with manipulation.

B.3 Using SPI to assess quality of counterfactual hires

In the paper, I estimate the relationship between supervisor characteristics and service delivery, construct SPI, and use SPI to assess the quality of counterfactual hires. However, there are a number of concerns with doing this. The key challenge with using this method to determine what supervisor characteristics are related to service delivery improvements is the possibility of endogenous matching of supervisors to clusters of CHWs. In particular, supervisors who are more educated or have higher Raven's scores (i.e. higher values of SPI) may oversee clusters of CHWs that were already on an upward performance trajectory, in which case the improvements are not attributable to the supervisor. Note that since this is a differences regressions, bias will result from differential trends rather than differences in levels. To test for differential trends, I use data from my baseline survey on the number of institutional deliveries assisted by each CHW in each of the six months prior to the supervisors beginning in their role (as well as the admin data for the period after supervisors began in their role). Appendix figure B.1 plots the coefficients from a regression of the number of deliveries assisted by CHWs in each month on supervisor SPI and finds no evidence of pre-trends, suggesting that improvements indeed reflect the supervisor.

A limitation of SPI is that it is only based on predictors that I am able to collect. It thus misses any unmeasured characteristics that are related to performance as a supervisor and are uncorrelated with the variables that I do collect. To get a sense of the significance of unmeasured characteristics, I take three alternative measures of quality of the hired supervisors and measure the extent to which they provide additional information on supervisor quality above and beyond SPI. These alternative measures of the quality of hires are: (1) performance evaluations of the supervisors by superiors; (2) the average of CHW ratings of their supervisors on a scale from 1 to 4 (collected during the second survey); and (3) a PCA index of process measures of supervisor performance.[31] Since the key outcome for policy is service delivery, I am interested in the extent to which they predict service delivery improvements beyond what is captured in SPI.

Appendix table A.3 regresses the change in CHW functionality score (the composite measure of service delivery) after the supervisor is hired (Δy_{ij}) on these measures as well as SPI. The performance evaluations and average CHW rating both predict improvements in service delivery, while the process measures do not (columns 2 to 4). However, only the CHW ratings have predictive power after SPI is added to the regression. This means that SPI is missing some of the information from CHW ratings about supervisor performance (column 5). However, the R^2 of a regression with just SPI is 0.162 (column 1), while including the additional three measures of supervisor quality only marginally increases R^2 to 0.187 (column 5). Thus this increased information about supervisor quality is potentially limited. Even though there are certainly aspects of supervisor quality aside from that captured by SPI, SPI appears to capture a meaningful fraction of observable variation.

Another issue is that the relationship between personal characteristics and performance could be different among the counterfactual hires. If corrupt individuals have a different production function than the non-corrupt, then using SPI to assess counterfactual hires could be misleading.

[31] The performance evaluations are the first principal component of nine ratings of the supervisors from individuals overseeing the program (discussed in more detail in appendix section B.2). The process measures come from the second survey of CHWs and are the frequency of supervisor interactions and number of supervisory tasks CHWS report their supervisor completing over the past two months.

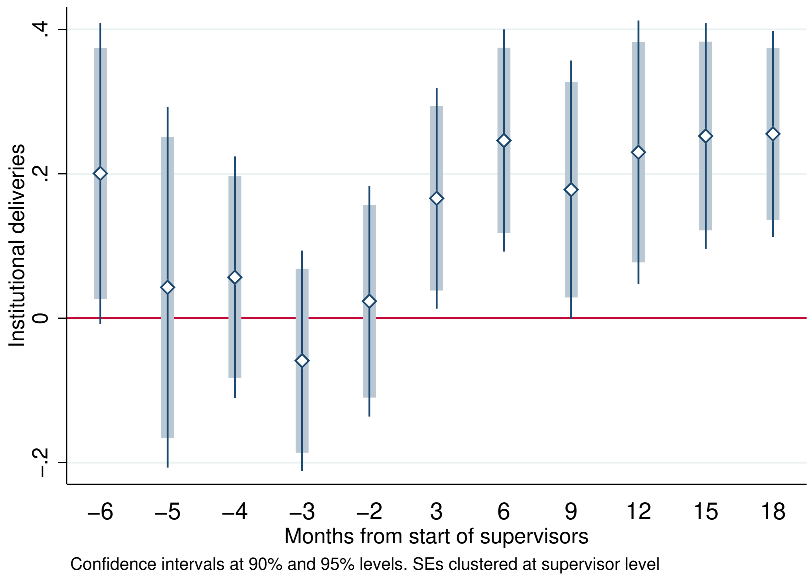
One encouraging metric is that the counterfactual hires are similar to actual hires in potentially corrupt actions. Taking the predicted counterfactual hires from the more rigorous endogenous entry model, over half admitted to offering a bribe for the job. Furthermore, nearly all of the predicted counterfactual hires also applied for the job in the actual hiring process (approximately 74% under the health knowledge test counterfactual and approximately 85% under the intended government system counterfactual). Given how well-known it was that paying a bribe was necessary to be hired, this suggests a willingness to engage in corruption.

We can also investigate this concern more empirically. Suppose that supervisors were randomly selected. The estimated relationship between individual characteristics and productivity as a supervisor from this randomly selected sample would generalize to the full set of CHWs. The key to whether the SPI relationship will extrapolate is whether there is heterogeneity in how supervisor characteristics are related to productivity along the dimensions on which supervisors are selected (e.g. education is relevant for productivity among those who offer bribes, but not for those who do not). I leverage the finding that hires were selected based on bribes, education, and political connections, and test whether these hiring-relevant factors interact with SPI in how supervisor SPI maps into changes in CHW performance (my main measure of supervisor performance). Appendix table B.3 finds no evidence of interactions, although the collinearity between SPI and education renders column (4) uninformative. While this test is not perfect (e.g. if the relevant dimension of heterogeneity is a binary measure of willingness to engage in corruption rather than the more continuous measure of bribe amount), it is suggestive. And given that most of the counterfactual hires were willing to engage in corruption, it is unlikely that this sort of heterogeneity could produce serious bias.

Another concern is that supervisors are hired out of the cluster of CHWs they oversee, but I use the change in performance of CHWs to measure the performance of supervisors. This could potentially generate bias where it might look like a cluster has better performance because the hiring process selected the CHW whose performance was growing most slowly; alternatively, a cluster may look like it is performing poorly because the hiring process selected the CHW whose performance was growing most quickly. As a check, I introduce the most extreme form of this bias and see how it would affect my results: I rank supervisors by their value of SPI and then add a duplicate observation to their cluster that is the opposite of their percentile ranking among supervisors (e.g. for the highest SPI supervisor, I duplicate the CHW with the worst change in performance in their cluster). I then reconstruct SPI in the new data, and redo the counterfactuals. After doing this, the estimated percent of first-best SPI for the actual supervisors is 90.3% (as compared to 90.6% in the main text), while in the counterfactuals is 85.5-89.4% under the intended government hiring system (86.1-90.5% in the main text) and 83.1-84.1% under the health knowledge system (82.3-84.2% in the main text). Thus it does not appear that this type of bias is economically meaningful. Intuitively, this is because supervisors typically oversee the work of 15-25 CHWs, and so removing just one from the pool of CHWs does not have much of an effect.[32]

[32]On the other hand, results change substantially when 3 or more CHWs are added under this method, but that is not what happened in the hiring.

Figure B.1: Test for Pre-Trends in Institutional Deliveries Related to Supervisor SPI



Notes: This figure plots the event study for the relationship between service delivery outcomes and the SPI of the hired supervisor. It plots the coefficients from a regression of the number of institutional deliveries assisted by a CHW on the interactions of whether the data is from a particular month and the SPI of the hired supervisor. This is really the combination of two event studies from different data sets: the months prior to the start of supervisors use data from surveys of CHWs, while the months following the start of supervisors use administrative data on deliveries. I combine the months following the start of supervisors into three month bins to reduce the amount of noise in the estimates. Confidence intervals are at the 90% and 95% levels. Standard errors are clustered at the cluster level.

Table B.3: Interactions Between Supervisor Performance Index and Selection Criterion

	Functionality Score	Functionality Score	Functionality Score	Functionality Score
Supervisor SPI	0.54	0.50	0.51	0.0077
	(0.059)	(0.226)	(0.061)	(0.407)
Supervisor SPI X Bribe		0.0029		
		(0.015)		
Supervisor SPI X Connection			0.11	
			(0.162)	
Supervisor SPI X Education				0.046
				(0.036)
Observations	917	906	917	917

Notes: This table studies whether there are interactions between the characteristics used to select supervisors and SPI in the production of improvements in CHW service delivery. This is used to study the validity of extrapolating SPI to non-hires. In all columns, the dependent variable is average monthly change in functionality score of the CHW after the supervisor is hired. Standard errors are clustered at the supervisor level.

C Supplemental Analysis

C.1 Further Investigation of the Hiring Process

The text of the paper uses a discrete choice model to study hiring decisions. I find that bribes appear to be the most important determinant of hiring, and quantitatively estimate the value that hiring agents place on political connections (4.68 salary-months of bribe) and education (each additional year of education is valued at 1.64 salary-months of bribe). This first part of this appendix takes a closer look at the correlates of bribe offers, while the second part considers a number of additional robustness checks and how to interpret the point estimates from the discrete choice estimation.

C.1.1 Bribe Offers

Bribe offers appear to be the primary driver of the selection decision, and so understanding what factors determine bribe offers can help explain the quality of hires. Table I explored some of these factors, but table C.1 goes into more detail. The first five columns of table C.1 use all of the reported bribe offers, while the second set of five columns replicates columns (1) to (5) but with only bribe offers from non-hires.

Columns (1) and (5) investigate two key relationships relevant to the main findings of the paper. As discussed in the text of paper using evidence from table I, wealth is strongly related to bribe offers – this is found in both columns (1) and (5) ($p = 0.001$ and $p = 0.004$). The coefficient on wealth in column (1) implies that bribe offers from the wealthiest applicants averaged around 6 months of salary more than the bribe offers of the poorest applicants. Given that the average *winning* bribe offer is seventeen months of salary, this helps explain the strong relationship between wealth and being hired in Table A.1.

I also find that individuals whose observable characteristics suggest they would perform better as supervisors (i.e. higher values of SPI, the index constructed in section III) offered larger bribes. The coefficient in column (1) implies that applicants at the 90th percentile of SPI offered bribes that averaged around five salary-months more than those at the 10th percentile. This supports the second explanation for why good candidates were selected under this corrupt system – better candidates were not only wealthier, but also had a higher valuation of the job independent of their wealth. As a result, they were willing to offer more money to be hired and thus were more likely to be selected.

That point is reinforced by the analysis in columns (2) and (6), which adds the variable "valuation proxy" to the regression. Valuation proxy is a measure of how much the candidate values the job, based on a question about whether they would be willing to apply under different probabilities of being selected.[33] If an individual is willing to apply at lower probabilities of selection, this implies they place a higher value on the job; for ease of interpretation, the probabilities are flipped to be equal to one minus the stated cutoff probability, so that 1 equates to the highest valuation. The coefficient on valuation proxy does not have a clear economic interpretation, but the sign and

[33] The question wording was: "Think back to the time when you were deciding to apply. Suppose that you knew the hiring process would occur without giving money or using connections. If you knew that *[X]* other women in your area were applying to become a supervisor, and you thought that you each had an equal chance, would you have applied?" The survey asked this for an ascending number of competitors.

statistical strength of the relationship are consistent with bribe offers being strongly related to willingness to pay. The inclusion of valuation also significantly attenuates the relationship between bribe offers and SPI, consistent with that relationship being due to a higher willingness to pay among the individuals better suited for the job. The relationship is not completely attenuated in column (2), but that is likely because the valuation proxy measure is relatively crude and so does not fully control for the SPI-valuation relationship.

Columns (3) and (8) examine whether bribe offers are responsive to the characteristics of the pool of individuals that the candidate is competing against. Candidates offer larger bribe offers when competing against a cluster of CHWs with a higher average value of the wealth variable. The "Average Wealth Bin" variable divides clusters into five quintiles, where the coefficient implies bribe offers averaged around four salary-months higher values in the wealthiest clusters than in the poorest. I also find that the bribe offers are higher in pools with a larger number of CHWs, where moving from a cluster at the 10th percentile to 90th percentile of number of potential competitors corresponds to an increase in bribe offers of nearly four salary months. The inclusion of these variables somewhat decreases the coefficient on wealth due to the intra-cluster correlation of wealth, but it remains strongly statistically significant (*p-value<0.05* in both columns).

These results are more consistent with an auction-style competition than alternative explanations, such as the hiring agents using a quality metric to make decisions and then demanding a bribe from the selected individual ("fully meritocratic hiring"). If the agent made selections based solely on skill, then the bribe price should not depend on the number of other potential applicants, but only the characteristics of the applicant related to their willingness to pay. Of course, the hiring agent may be making a threat that they would deviate to the next preferred candidate if the person selected will not pay the fee, and so the presence of competitors improves the threat level. But making that threat means the hiring agent is willing to trade off between bribe payments and quality (e.g. turning down a high quality candidate who only offers a bribe of zero), meaning that this is partially meritocratic hiring.

Columns (4) and (9) look at how some other characteristics of the candidate are related to the size of bribe offers. I do not find evidence of past performance as a CHW (as measured by the intended government criteria for hiring) or health knowledge being related to the magnitude of bribe offers, where those are the variables used in the two counterfactual exercises in the paper. Columns (5) and (10) examine robustness of those previous specifications to the inclusion of cluster fixed effects in order to account for unobserved heterogeneity. Average wealth and number of CHWs are dropped since they are fixed at the cluster level. There are also fewer observations since the fixed effects cause singleton observations to be dropped. Nonethless, the relationships between bribe offers and both wealth and valuation are statistically significant at the 10% level in all but one case (wealth in column (10), where the p-value is 0.122).

C.1.2 Characterizing Hiring Decisions

The paper argues that bribes, education, and connections are important determinants of hiring decisions. One way of gauging whether they are indeed important determinants of selection is to assess how well they predict the set of hires. For this, I use $\hat{u_{i,j}} = \hat{\theta}_1 b_{i,j} + \hat{\theta}_2 c_{i,j} + \hat{\theta}_3 \alpha_{i,j}$, the estimated utility of the hiring agent from hiring candidate i given the estimated coefficient values $\hat{\theta_1}$, $\hat{\theta_2}$, and $\hat{\theta_3}$. I calculate the percent of non-hires for whom $\hat{u}$ is lower than that of the winner of their competition (pairwise comparison). If there were no idiosyncratic element $\epsilon_{i,j}$ and the true parameters were known, then hires would always have higher values of u than a non-hire from their cluster. Bribes, connections, and education correctly predict the revealed preferred applicant in 88% of pairwise comparisons. Bribe offers appear to be the main driver of the selection decision, given that they alone correctly predict 82% of the pairwise comparisons, whereas education only predicts 56% and connections only predict 43%. Even if there are also other factors involved in hiring (or other factors that are correlated with these three factors), these explain a meaningful fraction of the hiring decisions.

I also carry out a number of robustness checks for the estimation. Some unsuccessful applicants refused to answer questions about bribes or claimed not to have offered a bribe. If these individuals did not offer bribes or would not have been competitive applicants, then their omission will not bias estimates. Parameter estimates depend mostly on individuals on the threshold of selection, since marginal changes in parameter values affect the probability that those individuals are selected. For those who are far from the threshold of being selected (either because they did not offer bribes or their offers were noncompetitive), changes in the parameter values have little effect on the overall likelihood expression, and so their inclusion makes little difference. This can be seen in column (1) of Table C.2, where non-hires who said that they did not offer a bribe are coded as offering a bribe value of zero (rather than a missing value). The estimates are virtually identical to those in Table I because these individuals are so far from the threshold that they do not affect the estimates.

However, it could be a problem if some unsuccessful applicants offered larger bribes than the selected applicants, but did not report them. This is unlikely for three reasons. First, unsuccessful applicants who did not report offering a bribe are poorer, at the 51.1^{st} percentile of wealth, than those who said they offered bribes (57.5^{th} percentile). Wealth is the most important determinant of bribe offers, so it is unlikely that they would have offered systematically larger bribes. Second, appendix table C.2 re-runs the estimation with imputed bribes based on observable characteristics for applicants with no reported bribe (using the statistically significant variables from table C.1). This has no effect on the estimates (column 2), and findings are robust to inflating the imputed bribe values considerably (column 3). Column (4) imputes bribes for anyone who did not self-report a bribe amount (including supervisors), and results are again similar. Third, I examine contests in which only a small fraction of candidates do not report bribe offers, i.e. the set of bribe offers is presumably more complete. Column (5) uses only competitions in which a third or fewer of the applicants do not have reported bribe offers, and findings are again similar.

A final concern is that applicant characteristics other than bribes, education, and connections may be relevant in the selection decision. I have tested other plausible predictors of hiring, including age, past management experience, wealth, past performance as a health worker, health knowledge, problem-solving abilities and psychometric measures, and do not find they are statistically significant predictors of hiring when bribes, education, and connections are included; wealth is however a strong predictor of hiring when those are not included, as seen in Table A.1.[34] There are certainly

[34] Column (6) of Table C.2 shows that wealth does not predict hiring decisions after accounting for bribes, education, and connections. The wealth-hiring relationship in Table A.1 appears to result from how wealth is related to bribes

many other unobservable qualities that I cannot observe, and so as a result, the values of the point estimates should be viewed as suggestive. However, selection on such unobservable characteristics would not change the main take-away, i.e. that the selection procedure is partially meritocratic. Even if education is correlated with an unobserved quality such as public spiritedness on which hiring agents actually make their decisions, that would still be meritocratic. Even if bribe amounts are partially correlated with unobserved quality measures that were selected upon, the correlation would have to be implausibly high for fully meritocratic selection to hold,[35] and this would go against the other pieces of evidence against fully meritocratic selection.

Another way to study correlates of hiring decisions is with a logistic regression rather than a discrete choice model. Under the latent variable interpretation of a logistic regression, the model would imply that there is a latent variable u that corresponds to the utility that the hiring agent gets from hiring a particular individual where $u = \beta_0 + \beta_1 bribe + \beta_2 education + \beta_3 connection + \epsilon$. The individual is hired if u is greater than zero where ϵ follows a logistic distribution. That model corresponds well to a setting in which applicants apply to a common set of jobs and the top candidates from the entire region are selected: the latent threshold corresponds to a cut-off above which individuals are selected. The discrete choice model more closely approximates the setting in this context, where rather than selecting all individuals with u greater than a threshold which is fixed across all the clusters, the agent compares among applicants within each cluster and selects the individual with the largest value of u. However, the logit model still can provide useful insights as an approximation of this setting, particularly with the inclusion of fixed effects to account for level differences across clusters.

Table C.3 remakes table I as a logit regression, where the average partial effects for the main variables of interest are reported at the bottom of the table. Again, bribe offers are robustly related to hiring decisions. Taking the average partial effects in column (1), these indicate that the average effect of a marginal increase in bribe offer of 1 salary-month is to increase the probability of selection by 4.7 percentage points. Similarly, the average effect of a marginal increase in education of the applicant by one year is to increase the probability of selection by 7.4 percentage points, while the average marginal effect of having a connection is to increase probability of selection by 22 percentage points. This translates into political connections being valued at approximately five salary-months of bribes, while a year of education is worth approximately 1.6 salary-months of bribes; these are basically the same as the discrete choice estimates in table I.

and education, both of which are selected upon by the hiring committee, rather than the hiring committee making decisions explicitly on wealth.

[35]For example, wealth is the best predictor of bribe size, but still only explains around a tenth of the variation in bribes; selection on unobservable factors would have to be implausibly stronger than on wealth to explain the selection on bribes.

Table C.1: Potential determinants of bribe offers

	All Bribe Offers					Non-Hire Bribe Offers				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Wealth	6.077	5.885	4.428	4.528	4.275	6.712	6.618	4.434	4.350	3.842
	(1.706)	(1.707)	(1.721)	(1.680)	(1.876)	(2.242)	(2.255)	(2.220)	(2.195)	(2.428)
SPI	0.727	0.489	0.342	0.537	0.296	0.433	0.238	0.0270	0.00114	-0.142
	(0.216)	(0.230)	(0.237)	(0.298)	(0.341)	(0.266)	(0.278)	(0.272)	(0.352)	(0.452)
Valuation Proxy		8.194	8.699	8.969	10.31		7.034	6.594	6.603	6.517
		(2.636)	(2.501)	(2.592)	(2.536)		(2.793)	(2.700)	(2.715)	(3.784)
Average Wealth Bin			0.877	0.884				1.256	1.264	
			(0.378)	(0.377)				(0.441)	(0.450)	
Number of CHWs			0.365	0.374				0.302	0.299	
			(0.117)	(0.117)				(0.140)	(0.143)	
Past Performance				-0.0303	0.128				0.0731	0.0174
				(0.246)	(0.281)				(0.331)	(0.419)
Health Knowledge				-3.037	-1.533				-0.172	-1.148
				(1.805)	(1.951)				(2.002)	(2.844)
Observations	189	189	189	189	177	124	124	124	124	107
Cluster FEs	No	No	No	No	Yes	No	No	No	No	Yes

Notes: This table regresses bribe offers (in months of salary) on potential determinants of bribes. Columns (1)-(5) use all reported offers, and columns (6)-(10) only use the offers reported by non-hires. Columns (1) and (6) examine how wealth and SPI are related to bribe offers. . Columns (2) and (7) add a measure of how much the individual values the job. Columns (3) and (8) examine how opponent characteristics affect bribe offers. Columns (4) and (9) test additional individual characteristics, and columns (5) and (10) add cluster fixed effects. There are fewer observations in columns (5) and (10) due to the inclusion of fixed effects resulting in singleton clusters being dropped. Wealth is the percentile rank of the individual in the wealth distribution of bidders. Average wealth bin splits the clusters into five quintiles based on average wealth of CHWs in the cluster, so the coefficient corresponds to the average increase in bribe value moving up one quintile in the wealth distribution. Total CHWs is the total number of CHWs in the cluster. Valuation proxy measures how much the candidate values the job, based on the minimum probability at which they would be willing to apply to the job. Health knowledge is their score on a test of health knowledge used in the counterfactuals. Standard errors are clustered at the supervisor cluster level.

Table C.2: Hiring decisions (robustness checks)

	Missing as Bribe of Zero	Non-Hire Imputed	Non-Hire Imputed x1.33	All Imputed	Missing <= 1/3	All
Bribe Amount	1.08	0.93	0.32	0.64	0.96	0.99
	(0.06)	(0.10)	(0.09)	(0.12)	(0.15)	(0.11)
Political Connections	4.90	6.44	5.65	5.91	3.79	4.73
	(1.16)	(1.12)	(1.15)	(1.14)	(1.67)	(1.21)
Education	1.56	1.16	1.14	1.16	1.60	1.56
	(0.24)	(0.24)	(0.24)	(0.24)	(0.42)	(0.29)
Wealth						0.49
						(2.28)
Observations	341	338	338	338	94	189

Notes: This table contains robustness checks for table I. Column (1) imputes a bribe value of zero for non-hires who say they did not offer a bribe. Columns (2)-(3) impute bribe offers based on observable characteristics for non-hires who say that they did not offer a bribe, with column (3) inflating that imputed value by 33%. Column (4) imputes bribes for both supervisors and non-hires who did not state a bribe amount. Columns (5) includes only competitions where there is potentially less missing data (fewer than a third of applicants). Coefficients are estimated via maximum likelihood and standard errors are based on the likelihood function.

Table C.3: Correlates of hiring decisions

	All	Primary Reports Only	Supervisor Imputed
Bribe offer	0.525	0.481	0.291
	(0.142)	(0.196)	(0.100)
Years of education	0.836	0.290	1.169
	(0.286)	(0.244)	(0.262)
Used connection (=1)	2.480	1.929	2.592
	(0.853)	(1.058)	(1.259)
Bribe APE	0.047	0.059	0.031
Education APE	0.074	0.035	0.123
Connection APE	0.220	0.235	0.272
Observations	191	89	102

Notes: This table reports the results of a logit regression of a binary indicator for whether an individual is hired as a supervisor on potential determinants of the hiring decision, such as bribes. All specifications include cluster-level fixed effects to account for unobserved heterogeneity at the cluster level. Bribe offer is their reported bribe offer (in months of supervisor salary), while connection is whether they reported using a connection to try to get the job. Column (1) includes all clusters. Column (2) includes only clusters in which I directly observe the bribe offer of the hire, while column (3) imputes the hire's bribe based on their observable characteristics in clusters in which the bribe offer of the hire is based on secondary reports from other CHWs. The table also reports the Average Partial Effect for the variables of interest in the main specification; since the samples differ, this is not necessarily comparable to the APE in other columns.

C.2 Counterfactual results with endogenous entry

Section III compared the actual hires to counterfactual hires under two merit-based hiring systems: a test of health knowledge, and the hiring rules that were supposed to be used. Since I do not observe these counterfactual hiring processes, this requires predictions of which individuals will apply for the job under the counterfactual hiring processes. In the paper, I made two simplifying assumptions on application behavior: (1) all who were interested in the job would apply; or (2) the set of applicants is the same as in the corrupt hiring process. While this simplifies the analysis, it may miss how the decision to apply under a particular merit-based system affects the quality of hires under that system. This appendix examines how endogenizing the application decision affects the results.

Determining the application decision

Economic theory states that a candidate will apply for a job if the expected benefit of applying exceeds the cost. The expected benefit is equal to how much the candidate values the job (v_i) multiplied by the probability of being hired if she applies. Under a merit-based system, no bribes are paid, and so the cost of applying, C_i, is the cost of attending an interview or taking a test. There will exist some cut-off probability e_i at which candidate i is indifferent between applying and not applying for the job ($e_i v_i = C_i$). The decision rule for applying is simple: a candidate will apply under a particular hiring system if her probability of being hired is greater than e_i.

I follow a four step process to endogenize application decisions. First, I use survey data and past application behavior to put bounds around e_i for each candidate i. Second, I estimate the probability of each candidate being hired under each of the counterfactual hiring regimes if they choose to apply, and third, I combine their cut-off probability and estimated probability of being hired to determine whether they would apply. Fourth, I examine the set of applicants for each position and determine which would be hired.

To build intuition prior to detailing the methodology, consider the example of one particular CHW in the data, which I denote CHW A. In the first step, I determine CHW A's cut-off probability e_i is bounded between 0.25 and 0.33. In the second step, since CHW A scored a 96% on the health knowledge test and there are 15 other CHWs in her cluster eligible to apply, her estimated probability of being hired if they apply as 0.71. This is higher than 0.33, so she will apply for the job. I follow the same procedure for all the individuals in their cluster and determine that three others would apply (CHWs B, E and G), with test scores of 83%, 85% and 91%. CHW A has the highest test score, so is the counterfactual hire.

Moving into the methodology, the first step uses survey data to place lower ($\underline{e}_i$) and upper ($\bar{e}_i$) bounds around e_i for each candidate. Using bounds is more robust to errors, so produces more credible estimates than trying to point identify e_i. During the survey, respondents were asked if they would apply for the job as supervisor under various probabilities of being selected.[36] The

[36]The question wording was: "Think back to the time when you were deciding to apply. Suppose that you knew the hiring process would occur without giving money or using connections. If you knew that *[X]* other women in your area were applying to become a supervisor, and you thought that you each had an equal chance, would you have applied?" If a respondent were willing to apply when facing two other women with equal chances, then this is a 1 in 3 chance), but not three other women with equal chances (1 in 4 chance), this implies that $.33 > e_i > .25$. The survey asked this for an ascending number of competitors.

bounds are taken from the point at which they are no longer willing to apply, e.g. if they are willing to apply at a probability of 0.33, but not at 0.25, then e_i must lie between those values (so $\underline{e}_i = 0.25$ and $\bar{e_i} = 0.33$).

Second, I estimate the probability of being hired for each candidate under each counterfactual. Selection under each counterfactual is based on some merit-based "score" such as a test score, where candidate i is hired if her score κ_i is the highest among the pool of individuals who applied for this job. I assume that each candidate knows her own score κ_i, but not the exact scores of individuals against whom she will compete. Instead, she knows the total number of individuals eligible to apply for the position, the distribution of scores in the population, $f(\kappa)$, and that her competitors have randomly and independently drawn their scores from a distribution $f(\kappa)$. Candidate i's probability of being hired is equal to the probability that all of the other candidates in her cluster either draw lower values of κ than her or prefer not to apply for the job.[37]

Application decisions will depend on a candidate's expectations of application strategies of other candidates, which makes the estimation of this probability challenging. To avoid those complications, I make a simplifying assumption: when a candidate considers the application behavior of individuals who draw higher values of κ than her, she assumes those individuals will apply as long as they "want" the job, i.e. their value of the job is higher than the cost of applying. Thus her probability of being hired is equal to the probability that no other candidate draws a higher κ than her and wants the job. Under this assumption, the probability of being hired is a simple function of a candidate's score κ and their number of opponents. I estimate this probability via simulation from the population of observed CHWs, generating 50,000 simulated clusters for each possible number of opponents by taking random draws with replacement from the full set of CHWs. For each of these simulated clusters, the "hired" candidate is the one with the highest score among those who want the job. The cumulative density function of the scores of simulated hires is the estimated probability that a candidate with that score from a cluster of that size gets hired. For example, a candidate with a health knowledge test score of 96% is estimated to have a probability of 71% of being hired in a cluster of 15 CHWs, but only 53% in a cluster of 25 CHWs. For each candidate, I estimate their probability of being selected under each counterfactual system given their score and the number of candidates eligible to apply for the same supervisor position.

The simplifying assumption means that candidates slightly underestimate their own probability of being selected, as there are some competitors with higher values of κ who would choose not to apply. In practice, this does not affect the main results because the object of interest is who is hired for the job. The only application decisions that matter for who is hired is that of people who are likely to be hired and by definition, these individuals have high probabilities of being hired. Thus anyone who has a higher score than them will have an even higher probability of being selected, and so is quite likely to apply if at all interested in the job. As a result, this simplifying assumption turns out to be a good approximation. Previous drafts of this paper developed a fully structural method that accounts for strategic expectations of other candidate's behavior. Results were almost

[37] Here I am assuming that the distribution of possible test scores is common knowledge. Another modeling option would be to assume that candidates know the realization of test scores for each of their potential opponents, but that there is uncertainty over their opponents' values of e_i (without this uncertainty, only the winner would ever apply). This seemed less intuitively plausible as a modeling assumption, where candidates are unlikely to know exactly how well one another would perform on a test or their exact performance as a health worker.

identical, indicating that this assumption is fairly innocuous.

Third, I determine the sets of potential applicants using the bounds and estimated probabilities of being hired. When considering their application decisions, candidates can be split into three sets. The first is *definite applicants*, whose probability of being hired is larger than the upper bound on their cutoff and so will definitely apply for the job. The second set is *definite non-applicants,* whose probability of being hired is below the lower bound on their cut-off and so will definitely not apply. For example, an individual with an upper bound of 0.33 and lower bound of 0.25 would be a definite applicant if their probability of being hired was 0.4 and a definite non-applicant if their probability of being hired was 0.05. The third set is *possible applicants*, whose estimated probability of being hired falls within their bounds. If the same hypothetical applicant had a probability of 0.29, they would be a *possible applicant,* since this probability falls between the upper bound of 0.33 and lower bound of 0.25, so it is not clear whether they would apply.

Finally, based on these three groupings, I determine the set of potential hires for each position, and based on that, place bounds around the SPI of the counterfactual hire. To determine the set of potential hires, I first take the individual with the highest score from the set of *definite applicants.* If all of the *possible applicants* have a lower score than that individual, then that individual is the hire. On the other hand, if some of the *possible applicants* have higher scores than that individual, it is possible that they would have applied and been hired. In that case, the set of potential hires would consist of the *definite applicant* with the highest score and any *possible applicants* with at least that score. The upper and lower bounds on SPI for that particular job will be the maximum and minimum values of SPI from that set of potential hires. These bounds end up being quite tight since the strongest candidates are typically willing to apply at low probabilities and so are *definite applicants.*

Counterfactual results

To assess the relative quality of actual and counterfactual hires, I use SPI as in the paper. As in the paper, the benchmark is the actual, corrupt hiring system, which produces hires at 90.6% of the predicted first-best SPI. Hires under the health knowledge test are at 83.5% to 84.2% of the first-best case. Figure III plots these bounds for the counterfactuals, along with bootstrapped 95% confidence intervals around both the lower and upper bounds. The upper bound is around 6 percentage points lower than first-best SPI under the actual corrupt system, and equivalence with actual hires can be rejected (lower bound *p-value = 0.001,* upper bound *p-value = 0.004).*

The second counterfactual examines who would have been hired under the intended government procedure discussed in section II.1, which is the closest to the true counterfactual. The percent of first-best SPI under the intended government method is bounded between 86.2% and 90.0%, slightly lower than that of actual hires (90.6%). While I can reject equivalence of the actual hires and the lower bound of hires under the government method *(p-value = 0.019),* I cannot reject equivalence with the upper bound of government hires: this indicates that outcomes would have been no better and perhaps slightly worse under the uncorrupted procedure.[38]

[38]This counterfactual omitted the interview score, which may be informative in selecting good supervisors. As a robustness check, I rerun the "best case" scenario for this counterfactual with interviews taken into account, in which interview scores are imputed as the candidate's SPI. The estimated percent of first-best SPI is still bounded between

Using the full endogenous entry model provides very similar results to those highlighted in the text of the paper, and in general, these results are robust to many different assumptions. Intuitively, this is because only the entry decisions of candidates with reasonably high probabilities of being hired in the counterfactuals end up mattering, as they are the ones who end up hired. Predicting the application behavior of those individuals is straightforward since almost all of the best candidates either would never apply or, more commonly, place a high value on this job. A high valuation means that they would be willing to apply even if their probability of being selected were relatively low. Since the best candidates by definition have a high probability of being selected, their estimated probability of selection is typically well above their application probability cutoff and so they choose to apply if they want the job.[39] Although this more complex procedure turns out to be unnecessary, it is comforting that the results are similar to those under simpler assumptions.

Although it is clear that there are differences between the actual hires and counterfactual hires under the health knowledge test, it is less intuitive how this translates into differences in service delivery. One way to make this translation is to do a back of the envelope calculation to express differences using the value of a statistical life. This also permits a comparison of the cost of potentially inefficient bribe payments to the gains from service delivery improvements by putting service delivery improvements in monetary terms. I conservatively focus only on the relationship between SPI and institutional deliveries since there is more extensive research on how institutional deliveries translate into lives saved. Combining the estimated relationship between supervisor SPI and institutional deliveries (appendix table A.2), a conservative estimate of the relationship between institutional deliveries and reduced neonatal mortality from a meta-review (Tura et al., 2013), and a conservative valuation of a statistical life ($100,000), having a supervisor with 0.1 SD higher value of SPI improves social welfare by a value of approximately 60 salary-months annually. The average differences in SPI between hires under the health knowledge test and intended hiring system are 0.40-0.44 SD and 0.01-0.23 SD respectively, which translates into meaningful differences in social welfare. Furthermore, when considering the costs and benefits of corruption, the value of hiring better supervisors far exceeds the value of bribe payments, suggesting that allocation of positions is the more important criterion to examine in evaluating counterfactual systems.

C.3 Auxiliary Tests on the Quality of Supervisors

One concern with the data used in the paper is that it was collected after hiring decisions had been announced. It is possible that supervisors may have gained skills after being hired and prior to data collection that complicate the comparisons between them and non-hires. This will not be an issue for characteristics that do not vary over this time frame (e.g. education, which all have completed, or cognitive ability). However, reading/writing ability, health knowledge, and performance as a CHW (in the period between their hiring and starting their work as supervisors)

88.9% to 90.6%, which does not change the overall conclusion.

[39] Thus even if the estimates of cutoff probability bounds or probability of being hired were slightly off, this would not affect the results. I test this by running a counterfactual where the perceived probability of selection is inflated to be 50% higher than I had estimated. The percent of first-best SPI for health knowledge counterfactual hires is 84.4-84.7% (as compared to 82.3-84.2% in the paper) and 87.1-90.7% under the intended government system (as compared to 86.2-90.5% in the paper). When I deflate their estimated probabilities by 50%, results are again similar: the range for the health knowledge test is 84.4-84.6%, while for the intended government system, the range is 86.0-89.8%.

may have responded to their hiring. I take two approaches to address this concern using data sources from before the hiring; this data is only available for a quarter of the study region, but still informative. The first to use the data to test whether there is indeed a differential change for supervisors relative to non-supervisors over the period between the supervisors being hired and my main data collection. The second is to test for relative quality of supervisors prior to hiring.

The first data set consists of tests of reading ability and health knowledge that were administered to CHWs approximately a year before the hiring of supervisors was announced, while the second is monthly-level data on CHW salaries beginning approximately a year and a half prior to the announcement of hiring of supervisors (see figure A.1 for a visual representation of the timing). Table C.4 uses the first data set to test for differential changes in reading ability and health knowledge for supervisors relative to the rest of the population. I use the two differences-in-differences specifications in equations 3 and 4, with data from the pre-hiring period ($t = 0$) and post-hiring data collection ($t = 1$). y_{ct} is the outcome for CHW c in period t, $post_{ct}$ is equal to a dummy for if the data comes from the post-hiring period, $supervisor_{ct}$ is a dummy for whether the CHW is hired as a supervisor, and ϕ_c is a CHW fixed effect. β_4 tests for differential changes over time for supervisors, and β_3 tests for differences between supervisors and non-hires in the pre-hiring period. For both health knowledge and reading test scores, I find that the supervisors had much better scores in the pre-period and no evidence of a differential improvement in the post period. This supports the use of post-hiring data for reading/writing ability and health knowledge.

$$y_{ct} = \beta_1 + \beta_2 post_{ct} + \beta_3 supervisor_{ct} + \beta_4 supervisorXpost_{ct} + \epsilon_{ct} \tag{3}$$

$$y_{ct} = \beta_1 + \beta_2 post_{ct} + \beta_4 supervisorXpost_{ct} + \phi_c + \epsilon_{ct} \tag{4}$$

The second approach examines the quality of supervisors relative to non-applicants and unsuccessful applicants in the pre-hiring period. I first look at salaries in the two years prior to hiring: CHWs are paid based on their delivery of specific health services in their village, and so the salary data reflects their performance in a given month. Column (1) of Table C.5 finds that the average earnings of CHWs selected as supervisors were 28% higher than those of unsuccessful applicants ($p = 0.029$), and 45.9% higher than those did not apply ($p = 0.007$) prior to the announcement of hiring for the supervisor position. This can be also seen in the top panel of figure C.1, which graphs a moving average of the salaries of non-applicants, unsuccessful applicants, and hires over time. Consistent with the data in the paper (table II), past performance of those selected as supervisors was better than that of those who were not selected.

Second, the earlier referenced health knowledge test was split into three types of questions – 12 questions on danger signs for mothers during pregnancy, 12 questions on danger signs for infants/mothers after delivery, and 18 questions on general health knowledge. The below table finds the same pattern as in table II: those who applied for the job as a supervisor had consistently better health knowledge (note that supervisors are among the set who applied for the job). In one of three categories, the difference between supervisors and unsuccessful applicants is statistically significant, so as in table II, it appears that supervisor's health knowledge is weakly better than unsuccessful applicants in the pre-hiring period.

	(1) Health Knowledge	(2) Health Knowledge	(3) Reading Ability	(4) Reading Ability
Supervisor	0.244		0.785	
	(0.0770)		(0.389)	
Supervisor X Post	0.0433	0.0433	0.314	0.314
	(0.109)	(0.0844)	(0.551)	(0.219)
Observations	474	474	474	474
Dependent Mean	0.489	0.489	4.006	4.006
CHW FEs	No	Yes	No	Yes

Table C.4: Differential Changes in Reading and Health Knowledge After Hiring

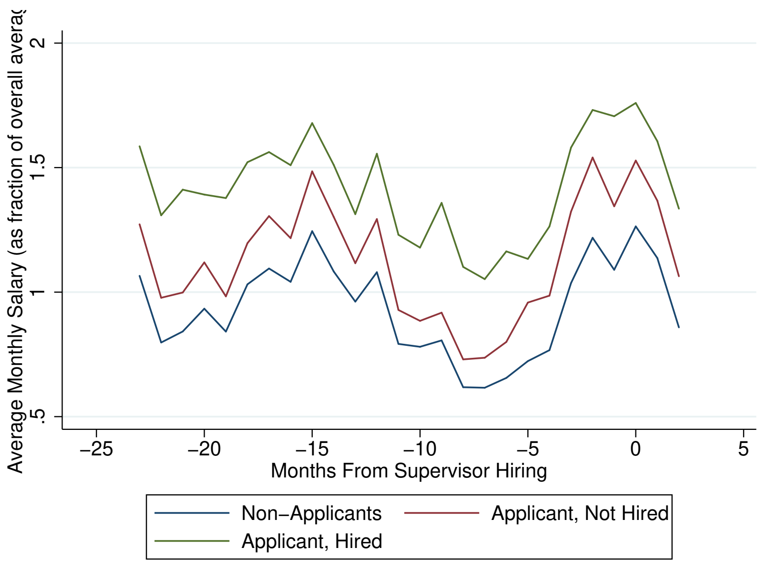
Notes: All specifications also include a dummy variable for whether the data come from the post-hiring period. Columns (1) and (2) test for differential changes in health knowledge for supervisors and non-supervisors, while columns (3) and (4) test for such changes in reading ability. Columns (2) and (4) include CHW fixed effects to improve the precision of the estimates. Standard errors are clustered at the CHW level.

Table C.5: Comparison Across Hires, Unsuccessful Applicants, and Non-Applicants

		Health Knowledge Test		
	Monthly Earnings	Pregnancy	Newborn Health	General Health
Hired (=1)	0.285	1.171	-0.214	-0.0571
	(0.130)	(0.404)	(0.212)	(0.429)
Applied (=1)	0.174	0.557	0.104	0.895
	(0.0639)	(0.200)	(0.105)	(0.211)
Dependent mean	1	8.7	10	15
Observations	236	236	236	236

Notes: This table reports estimates of the relationship between various measures of quality and whether a CHW was selected as a supervisor, using measures of quality collected prior to the hiring of supervisors.

Figure C.1: Salary of Hires, Unsuccessful Applicants, and Non-Applicants



Notes: This figure plots a moving average of the salaries of groups of CHWs before the CHWs hired as supervisors were selected. CHWs are paid on a piece rate, and so this reflects performance as a health worker.