Counterfactual thinking and causal mediation: An application to female labour force participation in India

Sunil Mitra Kumar* Ying-Fang Kao†

December 17, 2021

The use of computers has revolutionised our ability to learn about ourselves and the world around us. Beyond the goal of performance or prediction, the extent to which machine actions and algorithms are explainable and intelligible to human beings - Explainable AI - are increasingly becoming important, especially so in socio-economic contexts, and where life and health outcomes are involved. While the Turing test aims to distinguish between machine and human, Judea Pearl's 'mini' Turing test focuses on one crucial aspect of this distinction: the ability to reason causally and thereby answer causal queries based on counterfactuals. At the heart of counterfactual-based reasoning lies the role of causal explanation, or delineating the underlying causal mechanisms. In this chapter we make a small step towards demonstrating how causal models can be brought to observational data to answer useful counterfactual queries in contexts where complex social processes are at play. We estimate the causal effects of education on female labour-force participation in India in a causal mediation framework. We consider the role of positive assortative marital-matching in terms of education which leads to husbands' levels of education mediating the effect of women's education on their subsequent labour force participation, and we use a g-formula based approach to estimate the total causal and natural direct effect of education.

Keywords: Causal inference, causal mediation, female labour force participation

*India Institute and Department of International Development, King's College London, London, UK. Email: sunil.kumar@kcl.ac.uk

†Experimentation Team, Machine Learning and AI Division, Just Eat, London. Email: ying-fang.kao@just-eat.com

1 Introduction and motivation

The use of computers has revolutionised our ability to learn about ourselves and the world around us, and to solve problems, be they academic or practical. A range of approaches are now available to utilise computing machines to learn or infer knowledge from data, adapt and exhibit behaviour (often in complex environments), and together these can be termed *intelligent* in the broad sense of the term. These approaches, which are related in many ways and yet distinct in their aims, can be roughly classified into two broad categories: artificial and machine intelligence. The former involves developing computing machines geared towards understanding or mimicking human intelligence and cognitive capabilities. The latter approach, machine intelligence, is not strictly limited to human capabilities alone, and involves diverse applications across domains that make use of machine learning tools, for e.g. neural networks, classification algorithms, probabilistic learning methods, reinforcement learning and deep learning etc..[1] Although there is considerable debate on these matters, the role of an agent (human or artificial) - how they perceive information, adapt to their environment, decide and act in a goal-oriented manner - retains an important place in both traditions. High-performance, or maximising the chance of achieving a particular goal, often by learning from vast amounts of data, is no longer the absolute criteria to judge such algorithms. Rather, the extent to which these solutions or actions by machines are explainable and intelligible to human beings - Explainable AI - are increasingly becoming important too (Barredo Arrieta et al., 2020; Gunning et al., 2019). This is especially so in socio-economic contexts, and where life and health outcomes are involved.

An important link between the two traditions of AI is the Turing test (Turing, 1950) which has played a critical role in creating models or expert systems.[2] The Turing test was originally meant as a way to discern whether the machine or program in question can behave and learn just like human beings in a specific game setting. This can be useful in judging the level of performance, however, questions regarding the process by which such performance is achieved can be important too. While recent efforts in AI focus on building algorithms capable of human or expert level performance, it is still worth bearing in mind that human beings are capable of far deeper learning even with – compared to machines – extremely limited data and computational (cognitive) capacities (Kao and Venkatachalam, 2021). For instance, despite the success of deep learning in programs trained to play the game Go, there remains a significant gap between human capability and machine learning. In a similar vein, discussions around different versions of the Turing test acknowledge the gap between human and machine intelligence in socio-economics contexts (Pagliari, Bucciarelli and Chen, Pagliari et al.).

Judea Pearl, recipient of the Turing Award in 2011, proposes an interesting idea to test modern algorithms. Instead of a binary test like what Turing initially considered, Pearl proposes a Ladder of Causation, with progressively more advanced causal queries. Intellectual ability of a program or being can then be graded according to the level of causal reasoning. Pearl terms this the mini Turing Test (Pearl and Mackenzie, 2018). With the ability to observe association at the bottom level, and counterfactual thinking at the top, Pearl argues that even the most advanced deep learning algorithms in data science do not go beyond the first

[1] These twin categories are sometimes referred to as Classical and Modern AI (Russell and Norvig, 2009), or symbolic and connectivist approaches to AI.

[2] These are systems that are capable of yielding valuable insights and performing tasks as required once they have been trained with existing data.

rung of the Ladder. Given that counterfactual thinking - an ability to imagine and reason about potential worlds - is ubiquitous in everyday life and causal reasoning is an important arsenal in human intelligence, endowing algorithms with this ability becomes an exciting prospect.

There are several ways in which counterfactual thinking has been invoked in the literature: Chen and Du (2017) explore the link between counterfactual thinking in the context of generalised reinforcement learning and cognitive capacity of human subjects in relation to a beauty contest game. Learning patterns are tied to the notion of counterfactual thinking, since the latter are an important component in evaluating the payoffs associated with strategies under different, imaginary scenarios. Specifically, counterfactual thinking is characterised as in the attraction updating function, called the principle of simulated effect (Camerer and Ho, 1998). By including the principle of 'simulated effect', the experience-weighted attraction model generalises reinforcement learning and belief learning and allows for the individual agent to have varying degrees of counterfactual thinking-ability. Chen and Du (2017) are inspired largely by a bulk of psychological literature on associations between the level of counterfactual thinking and working-memory capacity. The primary hypothesis in this literature is that undertaking counterfactual thinking burdens the memory, which implies that heterogeneity in the level of counterfactual thinking could be explained by variations in the level of working-memory capacity. The experimental data from Chen's beauty contest game supports this hypothesis, and overall the study addresses the importance of counterfactual thinking in learning and decision-making in repeated economic games.

However, although not explicitly mentioned in the psychology and cognitive science literature, counterfactual thinking is not feasible without a map of causal pathways or structure. Pearl (2001, 2009) and Imai et al. (2010, 2011) provide a mathematical framework to formalise such structures using Directed Acyclic Graphs (DAGs), and discuss the identification conditions under which various causal, counterfactual-based queries can be answered using observational data. In this chapter our aim is to bring together some of the strands in these literatures, and to make a small step towards demonstrating how causal models can be brought to observational data to answer useful counterfactual queries. We do so in the context of the social sciences and economics. In this realm, answering causal queries remains a complex challenge, and the bulk of efforts focus on estimating treatment effects in contrast to delineating the underlying causal mechanisms. Even if we set aside the practical and ethical limitations involved in the use of randomised controlled trials with human beings in order to obtain unbiased causal estimates (Deaton and Cartwright, 2018; Imbens and Rubin, 2015), we note that such experiments are better suited to obtaining reduced-form estimates and but however are limited so far as understanding causal mechanisms go. There thus remains significant scope to develop methods that can help understand causal mechanisms.

Our demonstration focuses on an important socio-economic phenomena: the causal effects of education on female labour-force participation. As we discuss, the dominant approach in the literature on this topic focuses on establishing the causal relationship between education and labour force participation (LFP), and thus doing obviates the tougher problem of providing causal *explanation* for this relationship. Our attempt is to shed some light on the latter, and we study this problem in the context of India. Employing insights from the literature on marital matching, we propose a causal structure by which female education effects LFP not only directly, but also indirectly via the husband's level of education. Doing so, we hope to substantiate our claim that methods of causal inference designed to answer counterfactual-based queries in complex causal

systems can be productively put to use in the context of socioeconomic phenomena involving human beings and the highly complex social systems within which they exist.

1.1 Female labour force participation in India as a problem of causal mediation analysis

As economies develop, female LFP is believed to initially decrease and later increase, or in other words the association between the two phenomena ought to follow a U-shaped relationship (Klasen, 2019), even though empirical evidence suggests that there is significant divergence from this hypothesis at the level of individual countries depending on historical, cultural and economic context (Klasen et al., 2021; Jayachandran, 2021).[3] In India, a growing literature focuses on one aspect in particular of economic development: levels of education. There is significant evidence of a neutral-to-negative relationship between rising education and labour force participation amongst married women. Recent studies document that employment rates have remained stagnant (Klasen and Pieters, 2015) or declined (Afridi et al., 2018) alongside rising education levels and economic development overall, and it remains a puzzle why this is the case. Afridi et al. (2018) show that this fall has taken place only amongst married women in rural India, and suggest a potential explanation in the form of rising returns to home production relative to labour market participation for educated women. Mehrotra and Parida (2017) document similar trends but provide an alternative explanation. Structural transformation in agriculture has displaced females from this sector, while a combination of rising capital intensity in manufacturing and rising real wages in rural areas has resulted in limited alternative employment opportunities, while continuing patriarchal social norms imply limited flexibility all round.

However, average LFP over time is not to be confused with the cross-sectional relationship between LFP and education. As Datta Gupta et al. (2020) find, cross-sectional comparisons of LFP amongst married Indian women as a function of education levels do show a U-shaped relationship. Besides pointing out significant heterogeneity in this overall pattern, their explanations include rising education levels for nonmarket reasons – improved marriage prospects – limited employment opportunities for educated women, and unobserved self-selection processes. Chatterjee et al. (2018) also document a U-shaped relationship, and breaking this down by sector show that rising education correlates with rising salaried employment but declining participation in wage labour, family farms or businesses. As they discuss, in part this points towards the role of limited employment opportunities for educated females, but they also show that the higher household incomes apart from women's own earnings are associated with lower LFP for the women themselves.

The latter point is significant, because marriages in India are still overwhelmingly arranged even as women increasingly participate in spouse-selection with their parents (Allendorf and Pandian, 2016). This implies that social norms and specifically the intersection of caste, economic and educational status are still the primary forces that shape marital matches, and could therefore also play a significant role in explaining LFP for married women. As Lin et al. (2020) show, educational hypergamy has decreased in India in line with global trends, reflecting rising female education. However, while educational hypogamy is rising, women who marry down in terms of education are more likely to be marrying into a family with higher economic status than their natal family. The trade-off this evidences, the authors argue, suggests that economic resources

[3] See **?** also for a detailed overview of the challenges and controversies in defining female labour force participation.

play a larger role in marital matching than education levels with the result that women with higher education than their husbands are still likely to earn less than them. This finding is in consonance with Klasen and Pieters (2015)'s list of supply-side factors that depress women's LFP, including higher household incomes and husband's education.

Taken together, these findings have clear implications for better understanding the relationship between female education and LFP. First, that structural factors in the form of limited employment opportunities for educated females, particularly in rural areas likely play an important role. Second, that a combination of societal norms and income substitution effects at household level could be a significant part of the explanation. Third, establishing causality and delineating the underlying causal mechanisms is a consistent challenge throughout this literature. To our knowledge, there is limited evidence on both fronts. The main challenge in inferring causality is that several variables which help explain LFP and are also correlated with education levels are in fact intermediate variables or mediators, for example the education level of husbands and the income of the household as a whole. That is, these are factors causally effected by education and in turn causally effect labour force participation, and this causal structure ought to be explicitly taken into account to provide meaningful insights. Explaining the underlying causal mechanisms is even harder, since there are likely a combination of supply and demand-side factors in operation, and it is challenging to account for all of these simultaneously.

In this paper we address the first of these challenges and make some progress with the second. Specifically, we measure the causal effects of women's education levels on LFP in the cross-section, purposively accounting for the role of social norms within the household as one significant causal explanation that shapes this relationship. To do so we model the assortative matching between women and their husbands in terms of respective education levels, and the causal effects these education levels then have on women's eventual LFP once they are married. Doing so, we attempt to combine some of the insights Lin et al. (2020) provide for marital matching with their causal implications for the LFP trends documented by Chatterjee et al. (2018). The remainder of the paper is structured as follows. Section 2 presents the causal structure we propose to study this relationship, the causal estimands of interest, and estimation methods. Section 3 discusses the data used and section 4 presents our results. Section 5 concludes with a discussion of these results and the new insights offered.

2 Causal structure and methods

We would expect a woman's level of education to effect her husband's level of education via the process of assortative matching, most likely involving parents on both sides. As Lin et al. (2020) explain, this matching takes into account both education levels but also wealth levels of the respective families. Post-matching, we would then expect the husband's level of education to influence the wife's LFP through pathways including social norms and preferences as well as the husband's earning capacity. The resulting causal structure is portrayed as a Directed Acyclic Graph (DAG) in figure 1, where husband's education H mediates the effect of female's education E (treatment) on LFP (the outcome).

We focus on two causal estimands. For a given change in female education (E) from level e_0 to e_1, the Total Causal Effect (TCE) measures the resulting change in LFP accounting for both the direct effect of E

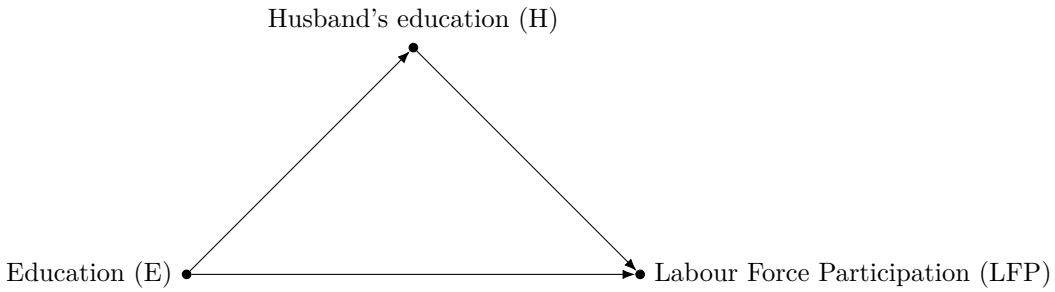


Figure 1: Causal structure for effect of women's education on labour force participation

as well as the indirect effect via the corresponding change in H. The Natural Direct Effect (NDE) focuses exclusively on one part of this: the change in LFP that is due to the change in E while holding mediator H fixed at the level it naturally attains at the baseline viz. $E = e_0$.[4] These twin causal quantities of interest are defined formally as follows, where $\text{LFP}_{eH_{e^*}}$ stands for labour force participation when female education is held at level e and husband's education H is held at the level it would attain when female education is at level e^*.

$$\text{TCE} = E[\text{LFP}_{e_1 H_{e_1}} - \text{LFP}_{e_0 H_{e_0}}] \tag{1}$$

$$\text{NDE} = E[\text{LFP}_{e_1 H_{e_0}} - \text{LFP}_{e_0 H_{e_0}}] \tag{2}$$

The above definitions apply when LFP is measured as a binary variable. We also consider LFP measured as a 4-category variable that captures the type of work undertaken (none, salaried, farm or business, wage labour). In this case, for each categorical outcome LFP_i where $i \in \{1, 2, 3, 4\}$ these estimands can be defined as:

$$\text{TCE}_i = \text{Prob}[\text{LFP}_{e_1 H_{e_1}} = \text{LFP}_i] - \text{Prob}[\text{LFP}_{e_0 H_{e_0}} = \text{LFP}_i] \tag{3}$$

$$\text{NDE}_i = \text{Prob}[\text{LFP}_{e_1 H_{e_0}} = \text{LFP}_i] - \text{Prob}[\text{LFP}_{e_0 H_{e_0}} = \text{LFP}_i] \tag{4}$$

2.1 Identification and estimation

Under the overall assumption that the data is generated from a Non-Parametric Structural Equation Model (Pearl, 2009, 2014), the following standard assumptions are required for identifying natural direct and natural indirect effects.[5]

A1: **Consistency**: (i) Consistency of E on H. (ii) Consistency of $\{E, H\}$ on LFP. This assumption requires that the actual and potential values coincide for all relevant variables.

[4] See Pearl (2001); Imai et al. (2011); Pearl (2014); VanderWeele (2015) for an introduction to causal mediation frameworks more generally.

[5] See VanderWeele (2015, Ch.2) and Pearl (2014) for a detailed discussion.

A2: **No unobserved mediator-treatment confounding**: There exists a set of variables W_1 such that $H_e \perp\!\!\!\perp E|W_1$. In other words, the effect of treatment on the mediator can be identified by conditioning on W_1.

A3: **No unobserved mediator-outcome confounding**: There exists a set of variables W_2 such that for each $e \in E$ we have that $\text{LFP}_{eH} \perp\!\!\!\perp H|\{W_1, W_2\}$. In other words, holding E fixed, the effect of the mediator H on the outcome can be identified. This also rules out any confounders of the mediator-outcome relationships themselves affected by E.

A4: **No unobserved treatment-outcome confounding**: There exists a set of variables W_3 such that $\text{LFP}_{eH} \perp\!\!\!\perp E|\{W_2, W_3\}$. In other words, holding H fixed, the effect of treatment on outcome can be identified by conditioning on $\{W_2, W_3\}$. This also rules out any confounders of the mediator-outcome relationship that are themselves affected by the treatment.

Estimating the causal effects listed above requires statistical methods that explicitly model the counterfactuals involved to keep mediators or treatment variables fixed while manipulating the variable of interest, since simply adjusting for the husband's education in a regression of LFP status on women's education and other covariates will in general not yield the required causal estimates.[6] Several methods are available for estimating direct and indirect effects in a causal mediation framework. The simplest of these entails specifying regression models for the mediator and outcome, and using the estimated parameters to provide analytical expressions for the expectation terms in equations 1-2. This can be challenging particularly when using non-linear regression models since it requires integrating over the conditional counterfactual distribution of the mediator. Monte-Carlo simulation can be used as an alternative, to directly estimate the required expressions through a two-step procedure. In step 1, using the same regression approach of one model each for the mediator and outcome, the estimated model parameters and the covariate values for each observation are then used to generate predictions for the mediator by holding the treatment at required levels. In step 2, using the regression estimates for the outcome model, covariates for each observation having replacing mediators with the predictions, and treatment held at required levels, are then used to generate predicted values for the outcome. Monte-Carlo simulation accounts for the sampling variability of the (conditional counterfactual) distributions of the mediator and outcome. We follow Daniel et al. (2015) and implement this by taking draws from the sampling distribution of the model predictions, obtaining standard errors via bootstrapping (Vansteelandt and Daniel, 2017; MacKinnon, 2008).

These steps are formalised as follows (see Daniel et al., 2015)

Let $\mathbf{W} = \{W_1, W_2, W_3\}$.

[6] In this type of regression, the causal interpretation of changing a representative female's education from one level to another in this sort of regression entails averaging over the sample distribution of all other included covariates, whereas the causal quantity of interest requires holding the husband's education fixed. In general, regression does not yield causal estimates if we condition on intermediate outcomes (see section 19.6 in Gelman et al., 2020) – here husband's education – unless we can assume that education does not interact with any other covariate in the model. But the latter is clearly unrealistic since we would expect the causal effect of interest to vary with the husband's level of education. Unfortunately, the simple cure of including an interaction term in the regression specification does not solve the problem either.

Step 1: Using OLS or maximum likelihood, estimate

$$E(H|E,\mathbf{W}) = g_h(A,\mathbf{W};\beta_\mathbf{h}) \text{ with error variance } \sigma_h^2$$
$$E(Y|E,H,\mathbf{W}) = g_y(E,H,\mathbf{W};\beta_\mathbf{y}) \text{ with error variance } \sigma_y^2$$

Step 2: For each unit i, holding $E = e, e^*$, draw

$$H_i(E,\mathbf{W_i}) \text{ from } N(g_h(E,\mathbf{W_i},\hat{\beta}_\mathbf{h}),\hat{\sigma_h^2})$$

Step 3: For each unit i, holding $E = e, e^*$ and $E' = e, e^*$, draw
$Y_i(E, H_i(E'))$ from $N(g_y(E, H_i(E'), \mathbf{W_i}; \beta_\mathbf{y}), \hat{\sigma_y^2})$.

Step 4: The empirical average across all units i of $Y_i(e^*, H(e))$, $Y_i(e, H(e))$ and $Y_i(e, H(e^*),)$ can be used to estimate the respective expectation terms in (1)-(**??**).

3 Data and variables

We use data from the second round of the India Human Development Survey (IHDS).[7] The IHDS is a nationally-representative household survey undertaken in two rounds, 2004-5 and 2011-12 (Desai et al., 2010; Desai and Vanneman, 2015), and contains a rich set of questions probing household and individual demographics and labour market participation. A supplementary questionnaire in the 2011-12 survey asks additional questions of a selected ('eligible') woman in a subset of the surveyed households. These questions cover topics including the nature of marital matching, such as the education levels of the woman's own parents and that of her in-laws, and whether the marriage was arranged.

As we now discuss, these covariates are essential for our analysis, and we therefore focus on the 2011-12 data, and within this, women interviewed as part of the 'eligible woman' supplementary questionnaire. We further restrict the analysis to women who have had an arranged marriage and who are in the age group 17-60 and therefore could potentially participate in the labour market. With these restrictions in place, our final sample size is 36,175.

In light of the discussion on causal structure above, an important part of the analysis is deciding which covariates to adjust for. Crucially, we want to avoid adjusting for intermediate outcomes unless they themselves are mediators such as husband's education, that is, unless they causally shape LFP. For this reason, we do not adjust for income or include it in our analysis. First, income is not observable in the dataset at the level of individuals other than for salaried employment. Second and more importantly, if household income is causally determined jointly by the woman's education and the husband's education and goes on to effect LFP, then it is an intermediate outcome and therefore should not be adjusted for. The same applies to household size and more specifically the number of children, since decisions about fertility are also, plausibly, effected by the husband's and wife's education levels jointly, and go on to effect LFP. An additional

[7] These are the same data used by Chatterjee et al. (2018)

complication with household size is that a woman might have caring responsibilities for children other than her own in the same household, and thus presents a potential missing data problem.

For these reasons, our analysis uses a smaller number of covariates than the covariates usually adjusted for in the literature. Specifically, our aim is to adjust for the minimal set of covariates that will plausibly fulfill assumptions A1-A4 presented above. In essence, these assumptions refer to identifying three types of causal effects: treatment on mediator, mediator on outcome holding treatment fixed, and treatment on outcome holding the mediator fixed. We now consider each of these in turn.

Identifying the effect of women's own education on that of their husband in fact refers to causal inference in the context of a historic event: the process of marital matching. Given that marriages are arranged, for any given level of the woman's education, her parents' own education (and thus social capital, expectations and aspirations) and income status ought to be sufficient to explain the choice of prospective husband and thus his education. Unfortunately, data on the parents' income are not available, however each woman is asked to list her mother and father's level of education. Of course the nature of marital matching might also vary according to religion and caste, and we therefore control for these as well, and additionally add interaction terms between the parents' education and caste group.

Next, to identify the effect of husband's education on female LFP, we want to adjust for any potential confounders that effect both variables. In the Indian context, this implies adjusting for variables that capture the social attitudes of the husband as well as attitudes of other household members whose views might shape the wife's LFP. To this end, our assumption is that the husband's parents would plausibly have influence over the woman's LFP, and therefore their education levels are a useful proxy for their social attitudes. As above, we recognise that this causal relationship likely varies across caste group and religion, and in the model we estimate below, we therefore include these variables together with their interaction terms with the husband's education. Analogous arguments apply for identifying the effect of the woman's education on her LFP while holding husband's education fixed. That is, beyond religion and caste group, we again assume that the in-laws' (i.e. husband's parents') levels of education are a useful proxy for the attitudes within the household that would help explain the woman's LFP.

Descriptive statistics for the variables used in our analysis are presented in table 1. We additionally also adjust for state (province) fixed effects, however these are not listed in the table for the sake of brevity. The two models described in Step 1 of section 2.1 can now be specified as follows

The mediator H (husband's education) is modelled using OLS where the vector $\mathbf{X_h}$ consists of education levels of the woman, her mother and her father, caste and religion dummies, interaction terms between caste-group and the woman's education, a dummy for urban area, and state fixed effects.

$$H = \beta_{h0} + \mathbf{X_h}\beta_{\mathbf{h}} + e_h \tag{5}$$

The outcome LFP is modelled using two separate specifications. For binary LFP we use a logit model where LFP_i represents the labour force participation of the i^{th} female and the vector $\mathbf{X_{LFP}}$ consists of a constant, education and its square, husband's education and its square, education level of the woman's in-laws, caste and religion dummies, interaction terms between caste-group and the husband's education as

well as the woman's education, a dummy for urban area, and state fixed effects.

$$\text{Prob}[\text{LFP}_i = 1|\mathbf{X_{iLFP}}] = \frac{\exp[\mathbf{X_{iLFP}}\boldsymbol{\beta_{LFP}}]}{1+\exp[\mathbf{X_{iLFP}}\boldsymbol{\beta_{LFP}}]} \tag{6}$$

For the categorical version of LFP that takes on four values $\{Y_1, Y_2, Y_3, Y_4\}$, we use a multinomial logit model with the same vector $\mathbf{X_y}$ as before. The vector $\boldsymbol{\beta_l}$ where $l \in \{1,2,3,4\}$ refers to the parameter estimates for outcome Y_l, and $\boldsymbol{\beta_1}$ is set to zero for identification.

$$\text{Prob}[\text{LFP}_i = Y_j|\mathbf{X_{iLFP}}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4}] = \frac{\exp[\mathbf{X_{iLFP}}\boldsymbol{\beta_j}]}{1+\sum_{l=2}^{4}\exp[\mathbf{X_{iLFP}}\boldsymbol{\beta_l}]} \qquad j \in \{1,2,3,4\} \tag{7}$$

4 Results

Table 2 presents the regression estimates for models (5), (6) and (7). The signs of the coefficients show that higher levels of education are associated with higher levels of husband's education, or in other words that there is positive assortative matching on this front, and the positive coefficient for the square of education shows that this relationship is convex. Education also has a convex relationship with LFP overall, but the negative coefficient for education and positive for its square suggest that the exact shape of the LFP-education relationship depends on the specific levels of education being considered. Things are more ambiguous when LFP is considered in terms of types of work and the relationship varies across the categories of work. Higher levels of husband's education however are associated with lower LFP throughout, while the positive coefficient for the square of husband's education indicates a convex relationship for overall (binary) LFP as well as salaried work. Taken together, these estimates suggests that the husband's education is positively effected by the woman's education, and it goes on to effect her LFP negatively, while the woman's own education has a convex association with her LFP that may or may not result in an overall positive association. However, on their own, these coefficients do not allow us to examine the U-shape hypothesis because of the two-step causal structure involved in determining this. Next, we therefore implement the steps provided in section 2.1 to estimate the total causal and natural direct effect of education on LFP.

The estimates of causal effects are presented graphically to aid interpretation. Figure 2 presents estimated Total and Natural Direct effect based on models (5) and (6), that is, when the outcome is binary LFP. The base level of education considered is 0 years, i.e. no formal schooling, since this is the mode of the distribution of education. The x-axis shows the levels of education at which causal effects are estimated. These correspond to the most common levels of education observed in the data which correspond to key stages: primary school (5 years), upper-primary (8), high school (10), higher-secondary (12) and graduate (15). These estimates are based on 200 bootstrap replications for each level of education shown, with 200 Monte-Carlo draws per estimate. Figure 2 supports the U-shaped hypothesis: comparing no formal education as the reference level (i.e. education=0) with primary and higher levels, LFP initially decreases, is lowest for upper-primary education, and then begins rising with higher levels of education (post-secondary education in particular).

Table 1: Descriptive statistics

N=36,175	Mean	S.D.	Min	Max
Labour market participation				
None	0.783	-	-	-
Salaried	0.038	-	-	-
Farm or business	0.061	-	-	-
Wage labour	0.118	-	-	-
Education (years)				
Own education	5.281	(4.933)	0.000	16.000
Mother's education	1.519	(3.079)	0.000	16.000
Father's education	3.509	(4.535)	0.000	16.000
Husband's education	7.127	(4.831)	0.000	16.000
Mother-in-law's education	1.148	(2.697)	0.000	16.000
Father-in-law's education	3.032	(4.263)	0.000	16.000
Age	35.779	(9.631)	17.000	60.000
Urban location	0.336	-	-	-
Religion				
Hindu	0.819	-	-	-
Muslim	0.121	-	-	-
Christian	0.022	-	-	-
Sikh	0.024	-	-	-
Buddhist	0.006	-	-	-
Jain	0.002	-	-	-
Tribal	0.004	-	-	-
Others	0.001	-	-	-
Caste group				
Brahmin	0.052	-	-	-
Forward/General (except Brahmin)	0.236	-	-	-
Other Backward Castes (OBC)	0.406	-	-	-
Scheduled Castes (SC)	0.212	-	-	-
Scheduled Tribes (ST)	0.081	-	-	-
Others	0.000	-	-	-

Notes:
This table presents means and proportions of all covariates used in the analysis. Survey probability weights are not taken into account in these statistics nor the analysis itself.

Table 2: Regression models

Dependent variable:	Husband's education	LFP (binary)	LFP by type of work (base: not in work)		
			Salaried	Farm or business	Wage labour
Education	0.379***	-0.084***	0.044	-0.014	0.095
	(0.022)	(0.025)	(0.045)	(0.040)	(0.067)
Education × Education	0.002*	0.014***	0.014***	0.001	-0.010***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Husband's education		-0.106***	-0.188***	-0.014	-0.102
		(0.029)	(0.047)	(0.045)	(0.062)
Husband's edu × Husband's edu		0.002***	0.006***	-0.001	-0.003*
		(0.001)	(0.001)	(0.001)	(0.001)
Mother's education	0.018*				
	(0.008)				
Father's education	0.128***				
	(0.006)				
Mother-in-laws's education		-0.019*	-0.010	-0.012	-0.098***
		(0.008)	(0.011)	(0.014)	(0.018)
Father-in-laws's education		-0.027***	-0.027**	-0.024**	-0.041***
		(0.005)	(0.008)	(0.008)	(0.008)
Caste group (Base category: Brahmin)					
Forward/General (non-Brahmin)	-1.616***	0.263	-0.467	0.353	0.421
	(0.181)	(0.229)	(0.430)	(0.358)	(0.443)
Other Backward Castes (OBC)	-2.061***	0.854***	-0.319	0.567	1.119**
	(0.173)	(0.223)	(0.411)	(0.350)	(0.433)
Scheduled Castes (SC)	-2.670***	1.230***	0.473	0.051	1.827***
	(0.177)	(0.225)	(0.413)	(0.358)	(0.434)
Scheduled Tribes (ST)	-3.189***	1.354***	-0.454	1.033**	1.849***
	(0.189)	(0.229)	(0.461)	(0.359)	(0.437)
Others	-2.067***	0.967***	-0.717	0.704	1.080*
	(0.319)	(0.294)	(0.806)	(0.453)	(0.493)
Education # Caste group (Base category: Brahmin)					
Forward/General (non-Brahmin)	0.106***	-0.069**	-0.030	-0.039	-0.102
	(0.019)	(0.023)	(0.041)	(0.037)	(0.066)
Other Backward Castes (OBC)	0.115***	-0.113***	-0.043	-0.056	-0.125
	(0.018)	(0.023)	(0.039)	(0.036)	(0.064)
Scheduled Castes (SC)	0.153***	-0.116***	-0.107**	-0.050	-0.137*
	(0.020)	(0.023)	(0.040)	(0.039)	(0.065)
Scheduled Tribes (ST)	0.213***	-0.090***	0.023	-0.078	-0.114
	(0.023)	(0.025)	(0.048)	(0.040)	(0.066)
Others	0.131***	-0.092*	-0.087	-0.038	-0.045
	(0.038)	(0.036)	(0.076)	(0.057)	(0.076)
Husband's education # Caste group (Base category: Brahmin)					
Forward/General (non-Brahmin)		0.033	0.062	0.006	0.043
		(0.027)	(0.042)	(0.043)	(0.062)
Other Backward Castes (OBC)		0.012	0.065	-0.015	0.043
		(0.027)	(0.041)	(0.041)	(0.060)
Scheduled Castes (SC)		0.008	0.094*	-0.015	0.045
		(0.027)	(0.042)	(0.043)	(0.060)
Scheduled Tribes (ST)		0.022	0.072	-0.005	0.053
		(0.028)	(0.049)	(0.044)	(0.061)
Others		-0.012	0.173*	-0.045	-0.007
		(0.039)	(0.080)	(0.061)	(0.072)
Urban	0.609***	-0.536***	0.348***	-0.966***	-0.790***
	(0.044)	(0.035)	(0.067)	(0.065)	(0.051)
Constant	6.566***	-1.860***	-3.027***	-3.113***	-4.075***
	(0.220)	(0.258)	(0.433)	(0.447)	(0.575)
Religion	yes	yes	yes	yes	yes
State (province)	yes	yes	yes	yes	yes
N	36321	36300		36300	

Notes: This table presents regression estimates for models (5), (6) and (7). Survey probability weights are ignored. Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

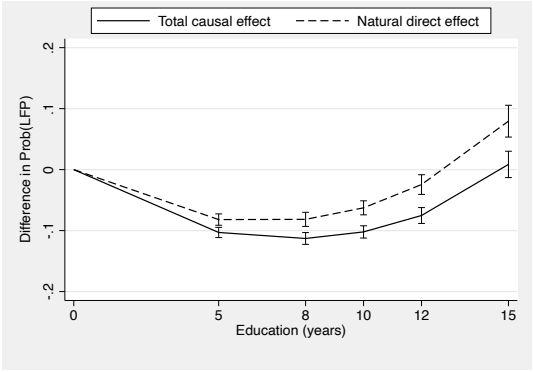
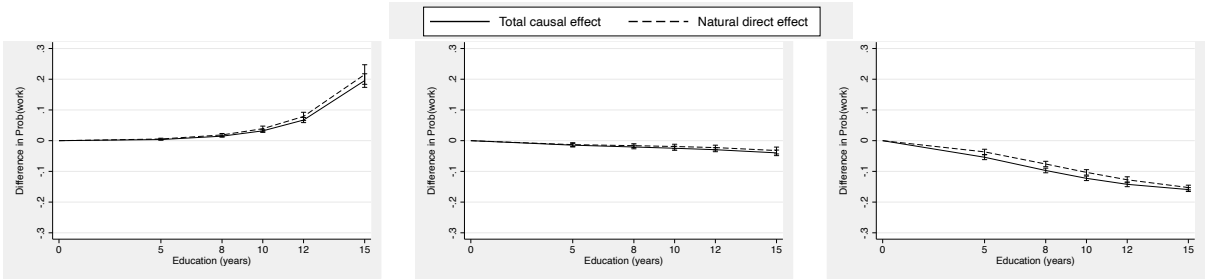


Figure 2: Causal effect of education on female labour force participation (reference level of education=0)

The position of the dotted line (NDE) above the solid line (TCE) shows that the direct causal effect of female education is more positive than its total causal effect. This makes sense, because the latter also takes into account the effect via husbands' education: the positive assortative matching due to marriage, and the negative effect of the husband's education on LFP, both of which factors combine to reduce LFP.

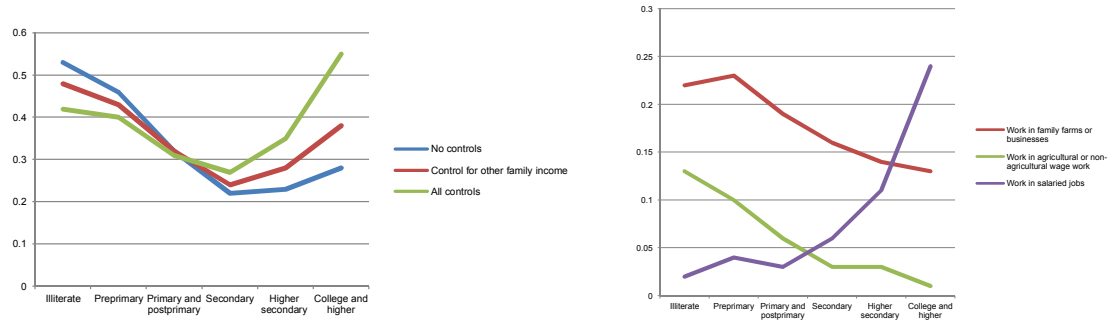
Figures 3a-3c present the estimates of total and natural direct effects according to type of work. These are based on models (5) and (7), that is, considering the categorical LFP outcome with three potential types of labour force participation. Other details are the same as those used for the estimates in figure 2, viz. the base level of education as 0 years, and estimates based on 200 bootstrap replications for each level of education shown, with 200 Monte-Carlo draws per estimate. These figures show that – as we might expect – the results for overall LFP mask significant heterogeneity depending on type of work. Across all three types of work, the TCE and NDE are very similar. For salaried work, the causal effect of education is positive and convex, with LFP rising slightly for primary, upper-primary and senior education, and rising significantly higher for senior-secondary and college education. For farm or business work the relationship is almost a straight line which slopes slightly downards, suggesting a near-linear, weakly negative relationship, while for wage labour the effects of education are nearly linear and more strongly negative. Recognising that wage labour accounts for the largest proportion of work, the U-shaped relationship for overall LFP therefore reflects the weighted average across these three types of work.



(a) Salaried work (b) Farm or business (c) Wage labour

Figure 3: Causal effects of education by type of work (reference level of education=0)

How do these results compare with the literature? The most direct comparison is with Chatterjee et al. (2018) whose main results are reproduced in figure 4 for ease of comparison. To reiterate, our results are based on a far smaller set of covariates for reasons discussed above, and, moreover, our estimation method is based on an account to explicitly model the causal structure involved. With these two significant differences in mind however, the results are very similar. The regression approach of Chatterjee et al. (2018) is most similar to estimating the NDE (albeit under stronger assumptions, including that husband's education is in effect measured pre-treatment). Comparing figure 4a with figure 2 shows that if we take the mediator-role of husband's education into account, the total causal effect of education is lower than the direct effect (in regression terms, the 'residual' effect), reflecting the negative role of (positive) assortative marital matching based on education combined with the negative effect of husband's education on women's LFP. To our mind this is a significant advantage of using causal mediation analysis, since it explicitly distinguishes between the two ways in which education effects LFP. Comparing figure 4b with figure 3 shows that while the sign and magnitude of effects by type of LFP are similar, our estimates are more attenuated.



(a) Overall LFP

(b) By type of work

Figure 4: Chatterjee et al. (2018)'s results: Predicted probabilities of labour force participation by education level
Source: Chatterjee et al. (2018) p.869 (overall LFP) and p.871 (LFP by type of work)

5 Discussion

Our motivation in this paper is to demonstrate how causal inference and counterfactual reasoning can be applied to disentangle complex social interactions using observational data. While counterfactual-based thinking is a well-studied psychological and cognitive phenomenon (Chen and Du, 2017) which has a critical role in advancing modern Artificial Intelligence (Pearl and Mackenzie, 2018) and many other domains besides including in the social sciences and economics, there is substantial variation in the attention paid to underlying causal mechanisms. In particular, the toolkits developed for causal mediation analysis in psychology and epidemiology are rarely deployed in economics to analyse social processes. By specifying the question of how education effects women's participation in the labour force in India as a problem of causal mediation, we have attempted to bring these literatures together, and demonstrate how causal mediation analysis can be a valuable tool in this domain.

We have argued why it is important to model education-based assortative marital matching as part of the overall causal explanation for how education effects labour force participation. Doing so explicitly recognises the role of the husband's education as a mediator via which female education shapes labour force participation. Similar to the literature, we find that the overall relationship between education and female labour force participation is U-shaped. However, we do so by considering the total causal effect and the natural direct effect separately. This lends the important insight that while both effects are U-shaped, the total effect is smaller because it also includes the negative effect of the husband's education – where the latter rises with rising female education due to positive assortative marital matching. We also consider the type of labour force participation across three categories, viz. wage labour, salaried work, and farm or business work. The effects of education vary significantly across the three types of work. Total and natural direct effects are very similar, with positive effects for salaried, negative for wage, and very slightly negative for farm or business work.

While the role of the husband's education as a mediator between (female) education and labour force participation is important to recognise, our causal structure nevertheless simplifies matters. For instance, it is plausible that the number of young children in the household also effects labour force participation, and is itself effected by education via fertility decisions and is thus an additional mediator. Further, fertility decisions might also be shaped by the husband's level of education level. Thus the number of young children might be a second mediator itself effected by husband's education as the first mediator. While this is just one example, in general, a more complex causal structure might be called for to study a phenomenon as complex as the labour force participation decision. Through our simple demonstration, our aim therefore is to argue for deploying causal mediation analysis as a valuable tool for investigating complex social processes, where reduced-form causal effects are indeed valuable, but far deeper insights can be gathered by understanding the underlying causal mechanisms.

References

Afridi, F., T. Dinkelman, and K. Mahajan (2018). Why are fewer married women joining the work force in rural India? a decomposition analysis over two decades. *Journal of Population Economics 31*(3), 783–818.

Allendorf, K. and R. K. Pandian (2016). The decline of arranged marriage? marital change and continuity in India. *Population and development review 42*(3), 435.

Barredo Arrieta et al. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion 58*, 82-115.

Camerer C. and T-H Ho (1998) Experience-Weighted Attraction Learning inCoordination Games: Probability Rules,Heterogeneity, and Time-Variation *Journal of Mathematical Psychology 42*, 305-326.

Chatterjee, E., S. Desai, and R. Vanneman (2018). Indian Paradox: Rising Education, Declining Women's Employment. *Demographic research 38*, 855–878.

Chen, S-H and Y-R Du (2017). Heterogeneity in generalized reinforcement learning and its relation to cognitive ability. *Cognitive Systems Research 42*, 1-22.

Daniel, R. M., B. L. D. Stavola, S. N. Cousens, and S. Vansteelandt (2015). Causal mediation analysis with multiple mediators. *Biometrics 71*(1), 1–14.

Datta Gupta, N., D. Nandy, and S. Siddhanta (2020). "Opt out" or kept out? the effect of stigma, structure, selection, and sector on the labor force participation of married women in India. *Review of Development Economics 24*(3), 927–948.

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine 210*, 2–21.

Desai, S. and R. Vanneman (2015). India Human Development Survey-II (IHDS-II), 2011-12. ICPSR36151-v2. Inter-university Consortium for Political and Social Research [distributor].

Desai, S., R. Vanneman, and National Council of Applied Economic Research, New Delhi (2010). India Human Development Survey (IHDS), 2005. ICPSR22626-v8. Inter-university Consortium for Political and Social Research [distributor].

Gelman, A., J. Hill, and A. Vehtari (2020). *Regression and Other Stories*. Cambridge University Press.

Gunning et al.(1998) (2019). XAI—Explainable artificial intelligence. *Science Robotics 4(37)*, eaay7120.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological methods 15*(4), 309.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*(4), 765–789.

Imbens G. and D. Rubin (2015) *Causal Inference: For Statistics, Social and Biomedical Sciences*. Cambridge University Press 2015.

Imbens, G. W. (2020, March). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *arXiv:1907.07271 [stat]*.

Jayachandran, S. (2021). Social norms as a barrier to women's employment in developing countries. *IMF Economic Review*, 1–20.

Kao, Y-F and R. Venkatachalam (2021). Human and Machine Learning. *Computational Economics 57*, 889-909

Klasen, S. (2019). What explains uneven female labor force participation levels and trends in developing countries? *The World Bank Research Observer 34*(2), 161–197.

Klasen, S., T. T. N. Le, J. Pieters, and M. Santos Silva (2021). What drives female labour force participation? comparable micro-level evidence from eight developing and emerging economies. *The Journal of Development Studies 57*(3), 417–442.

Klasen, S. and J. Pieters (2015). What explains the stagnation of female labor force participation in urban India? *The World Bank Economic Review 29*(3), 449–478.

Lin, Z., S. Desai, and F. Chen (2020). The Emergence of Educational Hypogamy in India. *Demography 57*(4), 1215–1240.

MacKinnon, D. (2008). Computer intensive methods for mediation models. *Introduction to statistical mediation analysis*, 325–346.

Mehrotra, S. and J. K. Parida (2017). Why is the labour force participation of women declining in India? *World Development 98*, 360–380.

Pagliari, C. , E. Bucciarelli , and S-H Chen (2021) Challenges in the Study of Intelligent Machines and Reverse Turing Test on Socio-Economic Decisions. *Decision Economics: Minds Machine, and their Society.* Edited by Bucciarelli, Chen, Corchado and Parra. Springer 2021.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality* (Second ed.). New York: Cambridge University Press.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods 19*(4), 459–481.

Pearl J. and D. Mackenzie (2018). *The Book of Why.* Penguin Random House

Russell, S., and P. Norvig (2009). *Artificial Intelligence: A Modern Approach (3rd Ed.)* Prentice-Hall, New York.

Turing, A. M. (1950) Computing machinery and intelligence. *Mind 59*(236), 433-460.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction.* New York: Oxford University Press.

Vansteelandt, S. and R. M. Daniel (2017, March). Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.) 28*(2), 258–265.