Online Appendix

Seeing Beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes

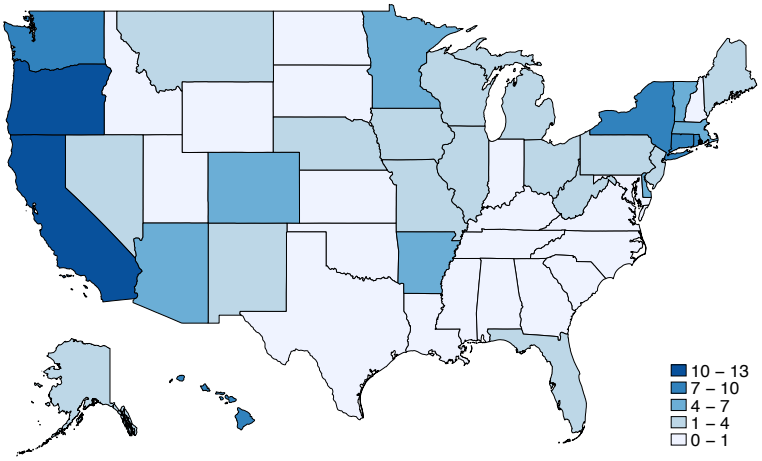
December 6, 2021

Doruk Cengiz	Arindrajit Dube	Attila Lindner	David Zentler-Munro
OM Partners	University of Massachusetts Amherst, NBER, IZA	University College London CEP, IFS, IZA, MTA-KTI	University of Essex

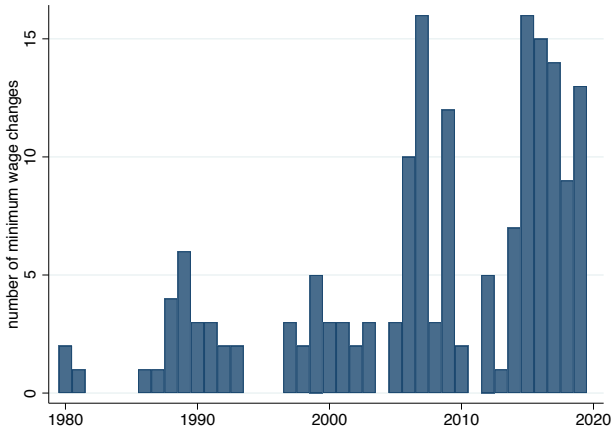
Contents

1 APPENDIX A: Additional Figures and Tables

Figure A.1: The Geographic Distribution and Timing of Prominent State level Minimum Wage Changes



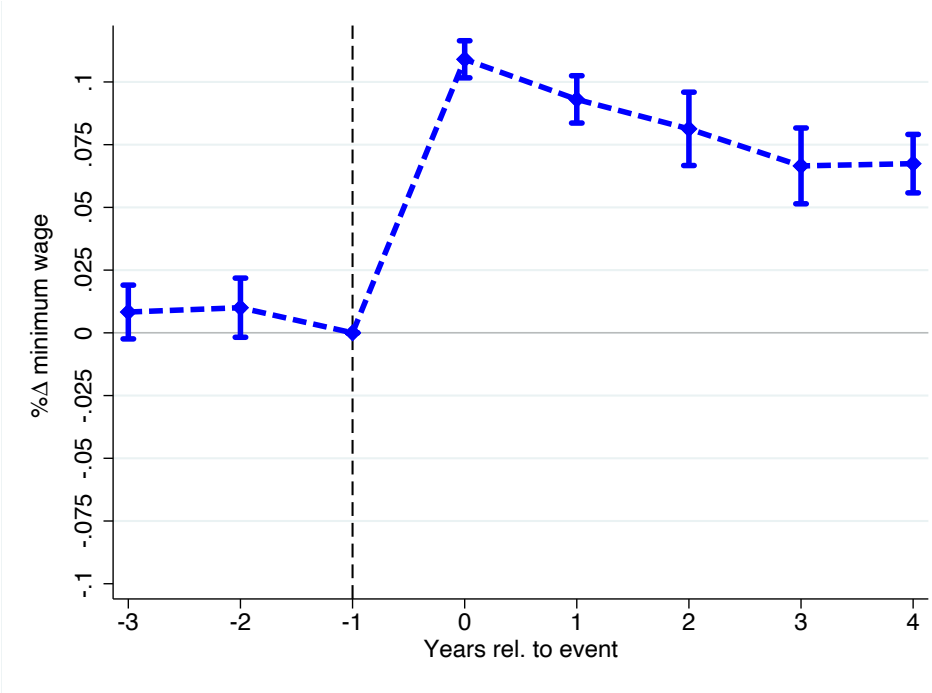
(a) Geographic Distribution



(b) Number of Events in Each Year

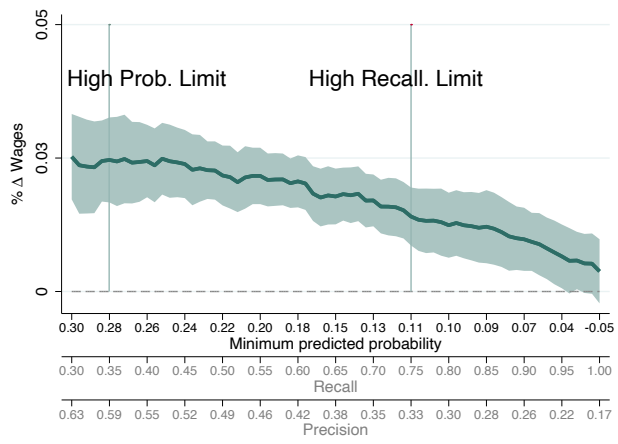
Notes: Panel (a) figure shows the geographic distribution of prominent state-level minimum wage changes occurred between 1979-2019. Panel (b) shows the number of prominent state-level minimum wage changes in each year. Prominent minimum wage changes are those where the (real) minimum wages increased by more than$0.25 and where at least 2 percent of the workforce earn between the new and the old minimum wage.

Figure A.2: The Evolution of Minimum Wages Following Prominent Minimum Wage Changes

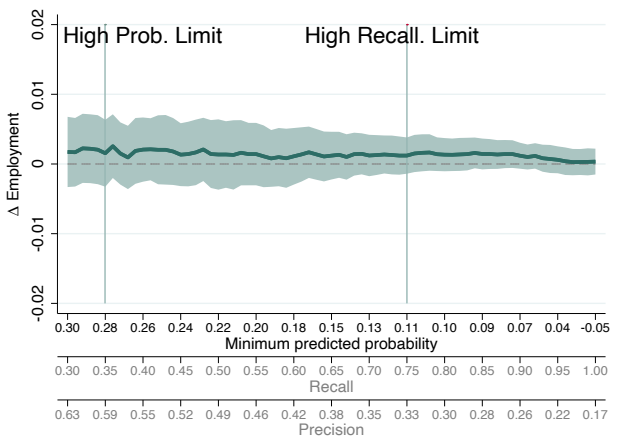


Notes: The figure shows the estimated change in minimum wages following a prominent minimum wage hikes. We apply our event study analysis (see equation 1) on (real) log minimum wages. We use 172 state-level minimum wage changes between 1979-2019. We also report the 95% confidence interval based on standard errors that are clustered at the state level.

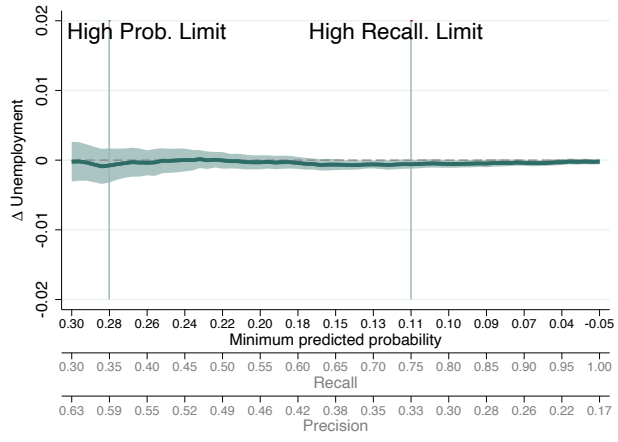
Figure A.3: Impact of the Minimum Wage for Alternative Predicted Probability Threshold Values, Card and Krueger Linear Prediction Model



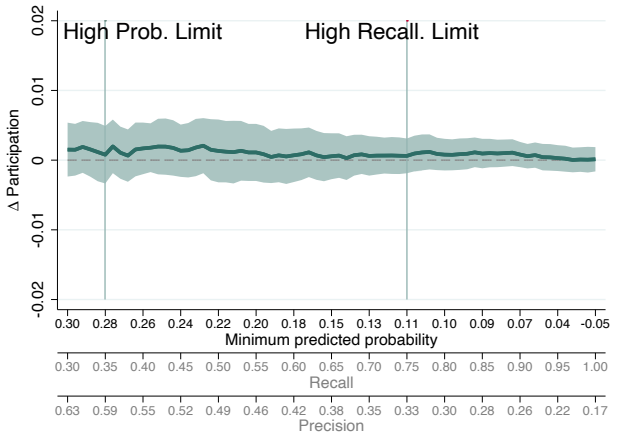
(a) Wage



(b) Employment



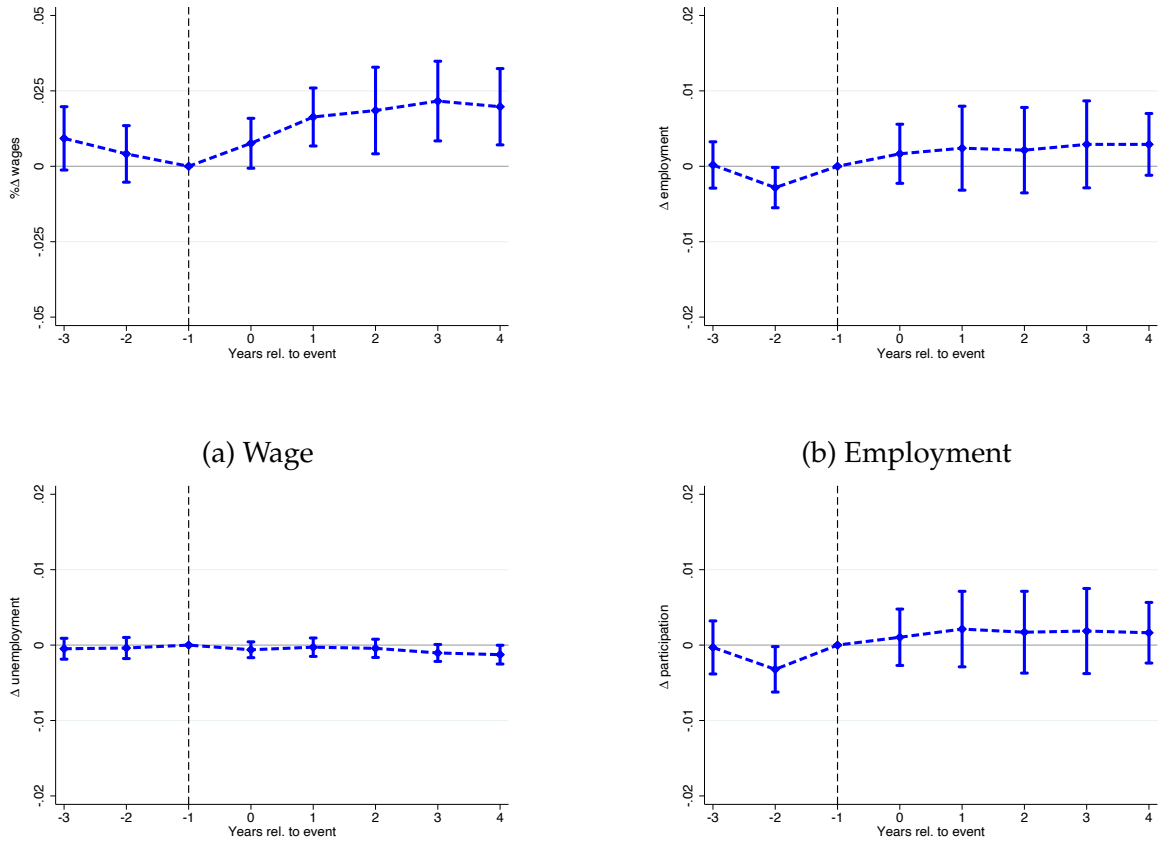
(c) Unemployment



(d) Participation

Notes: The figure shows the main results from our event study analysis (see equation 1) using alternative predicted probability threshold values. We exploit 172 state-level minimum wage changes between 1979-2019. Panel (a) shows the impact of the minimum wage on wages, Panel (b) on employment to population, Panel (c) on unemployment to population, and Panel (d) on participation rate. In each panel the green solid lines show the five year averaged post-treatment estimates for individuals whose predicted probability is above the "minimum predicted probability threshold". On the x-axis we also report the corresponding recall rate (the fraction of minimum wage workers retrieved by the prediction model if the particular minimum predicted probability threshold is applied) and the precision rate (the fraction of minimum wage workers in the sample if the particular minimum predicted probability threshold is applied). We also plot the thresholds corresponding to the high-probability group capturing 10% of the population with the highest predicted probability and to the high-recall group capturing 75% of all minimum wage workers. To calculate the predicted probabilities we use the Card and Krueger linear prediction model. The shaded areas show the 95% confidence interval based on standard errors that are clustered at the state level.

Figure A.4: Impact of the Minimum Wage Over Time (High-Recall Group), No Events After 2014q1



(a) Wage

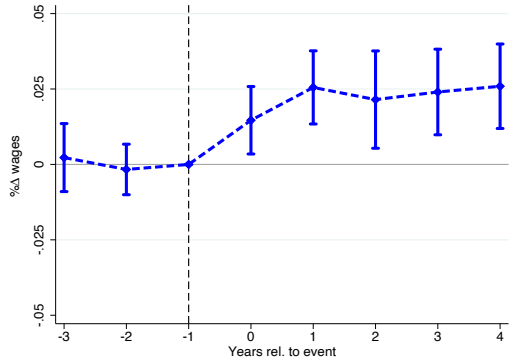
(b) Employment

(c) Unemployment

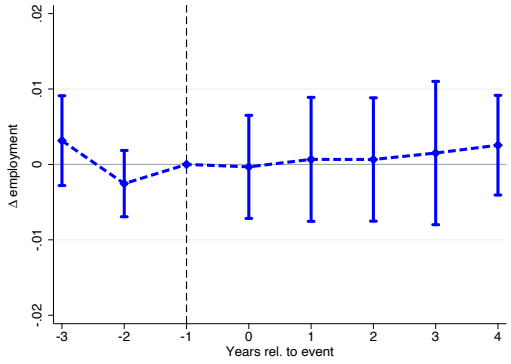
(d) Participation

Notes: The figure shows the main results from our event study analysis (see equation 1) using 99 state-level minimum wage changes between 1979-2014. While the estimation sample is between 1979-2019, here we exclude events for which we do not observe responses for the entire post-event window (5 years after). The figure shows the effect of the minimum wage increase on wages (Panel (a)), on employment to population (Panel (b)), on unemployment to population (Panel (c)) and on labor force participation rate (Panel (d)) for the high-recall group. The high-recall group consists of all workers whose predicted probability is above 11% – a threshold which corresponds to a 75% of recall rate. To calculate the predicted probabilities we use the best performing prediction model – the boosted tree prediction model. We also show the 95% confidence interval based on standard errors that are clustered at the state level.

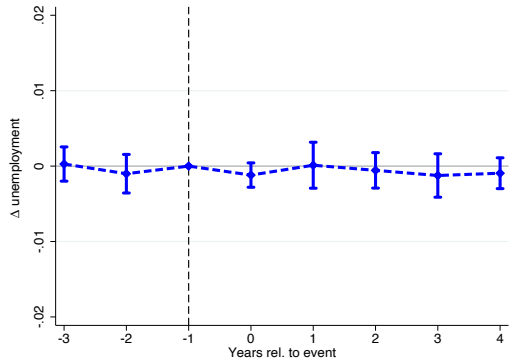
Figure A.5: Impact of the Minimum Wage Over Time (High-Probability Group), No Events After 2014q1



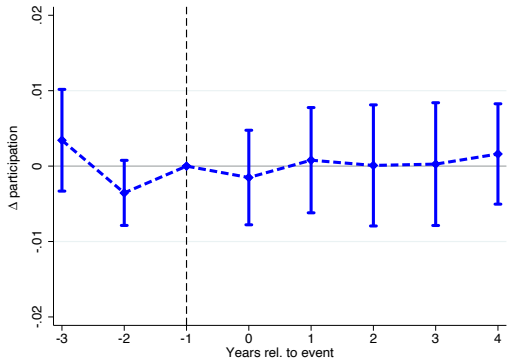
(a) Wage



(b) Employment



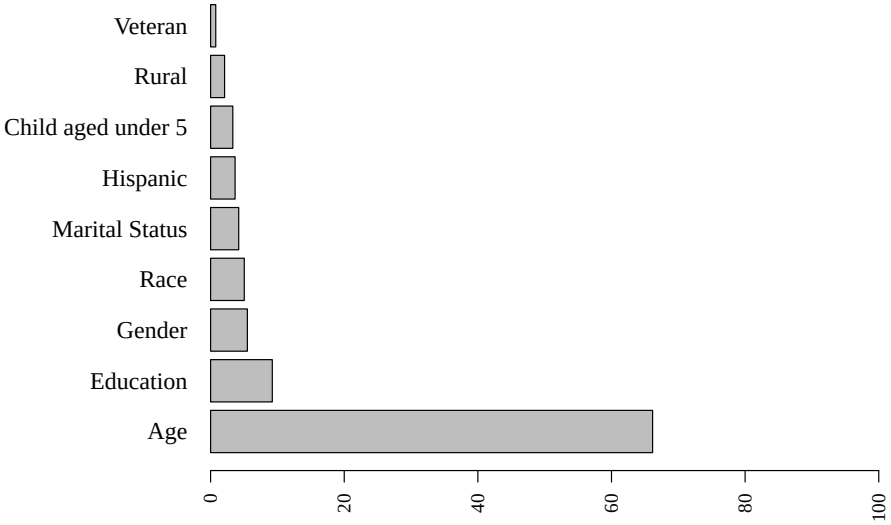
(c) Unemployment



(d) Participation

Notes: The figure shows the main results from our event study analysis (see equation 1) using 99 state-level minimum wage changes between 1979-2014. While the estimation sample is between 1979-2019, here we exclude events for which we do not observe responses for the entire post-event window (5 years after). The figure shows the effect of the minimum wage increase on wages (Panel (a)), on employment to population (Panel (b)), on unemployment to population (Panel (c)) and on labor force participation rate (Panel (d)) for the high-probability group. The high-probability group consist of 10% of the population with the highest likelihood of being affected by the policy. To calculate the predicted probabilities we use the boosted tree prediction model. We also show the 95% confidence interval based on standard errors that are clustered at the state level.

Figure A.6: Relative Influences of the Predictors in the Boosted Tree Prediction Model for Switching the Labor Force Status



Notes: We plot the relative influences of the variables in the boosted tree prediction model for switching the labor force status. We calculate the relative influence as in Friedman (2001) (see footnote 13 for the details). The bars, which indicate the decline in the loss function associated with the corresponding variable, are normalized so that they sum up to 100.

Table A.1: The Fraction of Minimum Wage Workers in Each Predicted Probability Decile

	Boosted Tree	CK Linear
Most likely decile	0.696	0.627
Probability decile 9	0.436	0.375
Probability decile 8	0.306	0.258
Probability decile 7	0.224	0.185
Probability decile 6	0.170	0.172
Probability decile 5	0.122	0.137
Probability decile 4	0.085	0.103
Probability decile 3	0.054	0.057
Probability decile 2	0.033	0.040
Least likely decile	0.020	0.022

Notes: The table shows the fraction of minimum wage workers at each predicted probability decile. Minimum wage workers are those workers earning less than 125% of the minimum wage. Column (1) applies the boosted tree prediction model, while Column (2) applies the Card and Krueger linear prediction model.

Table A.2: Predicted Probability Groups: Machine Learning and Linear Prediction

	Teen	20 $\leq$ Age <30	LTHS	HSG	Female	White	Black or Hispanic
High Recall: CK and ML	0.179	0.298	0.446	0.337	0.606	0.794	0.331
High Recall: ML not CK	0.000	0.056	0.000	0.863	0.704	0.754	0.262
High Recall: CK not ML	0.000	0.321	0.233	0.223	0.557	0.834	0.238
High Probability: CK and ML	0.704	0.212	0.672	0.178	0.629	0.799	0.377
High Probability: ML not CK	0.000	0.264	0.626	0.170	0.775	0.744	0.302
High Probability: CK not ML	0.124	0.505	0.694	0.224	0.453	0.872	0.640

Notes: This table shows the fraction of six samples of workers - one sample per row - that belong to the demographic group indicated in Columns 1-7. The samples are defined according to whether workers have a predicted probability, generated from the Card and Krueger (CK) linear prediction model or boosted-tree machine learning (ML) model, that put them in either the high recall or high probability groups. Row 1 shows workers that are classified as high-recall by both the CK and ML prediction models. Row 2 shows workers that are classified as high-recall by the ML model but not by the CK model. Row 3 shows workers classified as high-recall group by the CK model but not by the ML model. Row 4 shows workers that are classified as high-probability by both the CK and ML prediction models. Row 5 shows workers that are classified as high-probability by the ML model but not by the CK model. Row 6 shows workers classified as high-probability group by the CK model but not by the ML model.

Table A.3: Impact of the Minimum Wage on Labor Market Outcomes - Robustness to Alternative Specifications (High-Probability Group)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Δ wage (%)	0.021***	0.016***	0.022***	0.019***	0.022***	0.020***	0.020***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.003)	(0.003)
Δ employment (% pt)	0.001	0.005***	0.001	0.001	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)
Δ unemployment (% pt)	-0.001	-0.001	-0.001	-0.000	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Δ participation (% pt)	0.000	0.004**	0.000	0.000	0.000	0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
Employment Elas. w.r.t Min. Wage	0.035	0.159***	0.035	0.017	0.025	0.021	0.024
	(0.075)	(0.054)	(0.076)	(0.053)	(0.092)	(0.068)	(0.074)
Employment Elas. w.r.t Wage	0.159	0.923**	0.148	0.083	0.125	0.111	0.100
	(0.342)	(0.398)	(0.314)	(0.266)	(0.443)	(0.324)	(0.342)
Number of events	172	172	406	172	99	172	172
Number of observations	7,854	7,854	7,854	7,854	7,854	7,854	7,854
Number of individuals in the sample	7,581,999	7,581,999	7,581,999	7,581,999	7,581,999	7,581,999	7,581,999
Mean employment	0.347	0.347	0.355	0.360	0.363	0.347	0.347
Mean unemployment	0.057	0.057	0.062	0.056	0.063	0.057	0.057
Mean participation	0.404	0.404	0.417	0.416	0.426	0.404	0.404
Controls:							
State FE	Y	Y	Y	Y	Y	Y	Y
Quarter FE	Y	Y	Y	Y	Y	Y	Y
Division-Quarter FE		Y					
State Federal Events			Y				
Unweighted				Y			
No Events After 2014q1					Y		
State Employment Control: All						Y	
State Unemployment Control: All						Y	
State Employment Control: Low Prob. Group							Y
State Unemployment Control: Low Prob. Group							Y

Notes. The table reports the effects of the minimum wage on labor market outcomes based on the event study analysis (see equation 1) using 172 minimum wage changes between 1979 and 2019. We assess the impact of the minimum wage on the high-probability group. The high-probability group consists of 10% of the overall population with the highest likelihood of being affected by the policy. The table reports five year averaged post-treatment estimates for each key labor market outcome: percent change in wages and the change in employment to population, unemployment to population, and labor force participation rate. We also report the employment elasticity with respect to the minimum wage and the employment elasticity with respect to the wage, which is the ratio of the percent change in employment and wage. To calculate the percent change in employment we divide the change in employment to population by the mean employment to population rate preceding the minimum wage hikes (reported at the bottom of the table). The line on the number of observations shows the number of quarter-state cells used for estimation, while the number of individuals refers to the underlying CPS sample used to calculate labor market outcomes in these cells. In all the regressions we use the best performing prediction model — the boosted tree prediction model. The first column shows the preferred benchmark estimate reported in Column (1) of Table 2. Column (2) augments the baseline model with division-by-quarter fixed effects . The third column reports estimates using 406 state and federal minimum wage increases. All regressions are weighted by state-quarter population except Column (4), where we report unweighted estimates. Column (5) only considers minimum wage events that happened on or before 2014q1 to ensure a full five year post-treatment period. Column (6) controls for overall state-level unemployment and employment rates (as a fraction of population), while Column (7) controls for the employment and unemployment rates of individuals with low predicted probability of being a minimum wage worker (less than 11%). Robust standard errors in parentheses are clustered by state; significance levels are * 0.10, ** 0.05, *** 0.01.

Table A.4: Impact of the Minimum Wage on Labor Market Outcomes: Quintile Analysis

	(1)	(2)	(3)	(4)	(5)
Δ wage (%)	0.022***	0.008**	0.001	-0.000	-0.003
	(0.003)	(0.004)	(0.005)	(0.003)	(0.004)
Δ employment (% pt)	0.001	0.001	0.001	0.001	-0.001
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Δ unemployment (% pt)	-0.000	-0.001**	0.000	0.000	-0.001*
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Δ participation (% pt)	0.001	0.000	0.001	0.001	-0.002*
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Employment Elas. w.r.t Min. Wage	0.036	0.027	0.010	0.013	-0.016
	(0.055)	(0.026)	(0.026)	(0.009)	(0.014)
Employment Elas. w.r.t Wage	0.153	0.337	0.927	-12.123	0.430
	(0.227)	(0.410)	(3.754)	(416.674)	(0.394)
Number of events	172	172	172	172	172
Number of observations	7,854	7,854	7,854	7,854	7,854
Number of individuals in the sample	14,382,008	11,318,603	9,138,322	8,030,534	7,418,458
Mean employment	0.379	0.521	0.672	0.789	0.863
Mean unemployment	0.050	0.037	0.036	0.028	0.021
Mean participation	0.429	0.557	0.708	0.817	0.884
Group	Fifth Quintile	Fourth Quintile	Third Quintile	Second Quintile	First Quintile
Prediction Model	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree

Notes: The table reports the effects of the minimum wage on labor market outcomes based on the event study analysis (see equation 1) using 172 state-level minimum wage changes between 1979 and 2019. The table reports five year averaged post-treatment estimates for each key labor market outcome: percent change in wages and the change in employment to population, unemployment to population, and labor force participation rate. We also report the employment elasticity with respect to the minimum wage and the employment elasticity with respect to the wage, which is the ratio of the percent change in employment and wage. To calculate the percent change in employment we divide the change in employment to population by the mean employment to population rate preceding the minimum wage hikes (reported at the bottom of the table). The line on the number of observations shows the number of quarter-state cells used for estimation, while the number of individuals refers to the underlying CPS sample used to calculate labor market outcomes in these cells. Columns (1) to (5) show estimates for the top through to bottom quintile respectively. Robust standard errors in parentheses are clustered by state; significance levels are * 0.10, ** 0.05, *** 0.01.

Table A.5: The Impact of the Minimum Wage on Single Mothers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Δ wage (%)	0.006*	0.006*	0.013***	0.013***	0.013***	0.026*	0.026*	0.028**	0.028**	0.028**
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.014)	(0.014)	(0.011)	(0.011)	(0.011)
Δ employment (% pt)	0.001	0.000	0.001	0.001		0.002	0.010**	0.002	0.007	
	(0.001)	(0.001)	(0.001)	(0.002)		(0.004)	(0.004)	(0.005)	(0.008)	
Δ unemployment (% pt)										
Δ participation (% pt)	0.000	0.000	0.001	0.000	0.000	0.001	0.007	0.001	0.004	0.004
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.004)	(0.006)	(0.006)	(0.009)	(0.009)
Employment Elas. w.r.t Min. Wage	0.009	0.002	0.025	0.025	-0.004	0.029	0.178**	0.037	0.135	0.382**
	(0.017)	(0.020)	(0.031)	(0.038)	(0.060)	(0.064)	(0.078)	(0.103)	(0.148)	(0.182)
Employment Elas. w.r.t Wage	0.153	0.025	0.184	0.182	-0.017	0.106	0.655	0.129	0.473	0.907
	(0.238)	(0.323)	(0.202)	(0.260)	(0.271)	(0.239)	(0.491)	(0.372)	(0.632)	(0.716)
Number of events	156	156	156	156	156	156	156	156	156	156
Number of observations	6,222	6,222	6,222	6,222	6,222	6,222	6,222	6,222	6,221	6,221
Number of individuals in the sample	39,257,290	9,851,316	17,343,511	4,347,519	4,347,519	672,647	169,347	559,702	140,903	140,903
Mean employment	0.613	0.611	0.432	0.430	0.261	0.568	0.563	0.528	0.524	0.356
Mean unemployment	0.037	0.047	0.046	0.060	0.060	0.083	0.107	0.090	0.116	0.116
Mean participation	0.650	0.658	0.479	0.489	0.489	0.651	0.670	0.618	0.640	0.640
Prob. Group	All	All	High Recall	High Recall	High Recall	All	All	High Recall	High Recall	High Recall
Employment dataset	Basic	MORG	Basic	MORG	MORG	Basic	MORG	Basic	MORG	MORG
Employment measure	Emp. Status	Emp. Status	Emp. Status	Emp. Status	$w > 0$	Emp. Status	Emp. Status	Emp. Status	Emp. Status	$w > 0$
Demog. Group	All	All	All	All	All	Single mother kids under 5	Single mother kids under 5	Single mother kids under 5	Single mother kids under 5	Single mother kids under 5

Notes: The table compares the impact of the minimum wage on the entire population of workers (Columns 1-5) to the impact of single mothers with children under 5 (columns 6-10). Columns 1 and 2 use data from the CPS basic monthly sample and monthly outgoing rotation group (MORG) sample respectively to estimate employment effects for the entire population of workers. Columns 2 and 3 do likewise for the high recall probability group of workers. Column 5 changes the measure of employment from the self-reported employment status to reporting a positive wage, which can only be done for the MORG sample. Columns 6-10 follow the same format as Columns 1-5 for single mothers with children aged under 5 rather than the entire population of workers.

2 APPENDIX B: Data Sources and Variable Construction

The data sets in the main text are as follows:

We use the 1979-2019 CPS-Outgoing Rotation Group (CPS-ORG) in building the prediction model, and in estimating the wage effects of the minimum wage. In constructing the hourly wage variable, we exclude self-employed workers as well as observations with imputed wage information. Following Feenberg and Roth (2007)'s recommendation, if the individual is not paid hourly, we calculate the hourly wage by dividing earnings per week by usual weekly hours worked in the job. We download this dataset from the website of the National Bureau of Economic Research (NBER).

To obtain state-by-quarter labor force statistics (employment, unemployment, participation, self-employment and part-time/over-time status) as well as information on number and age of children, we rely on the 1979-2019 Basic Monthly CPS dataset downloaded from the Integrated Public Use Microdata Series (IPUMS) website: for further details see Flood et al. (2020).

We use state-level quarterly statutory minimum wage data from Vaghul and Zipperer (2016), updated by the authors through 2019.

In building the prediction models, we use the following variables:

Minimum wage worker indicator: The outcome variable that takes on the value of 1 if the individual's hourly wage is less than 125% of the statutory minimum wage. We also do not include self-employed workers or those with imputed wages.

Age: We use reported age throughout the text.

Education: We construct a categorical variable with four categories, less than high school (EDUC=1), high school graduate with no college education (EDUC=2), some college (EDUC=3), and college graduate (EDUC=4), using the variables that report highest degree completed.

Gender: We construct a binary variable that takes on the value of 1 if the individual is male, and 0 otherwise.

Rural residency: We construct a categorical variable with two categories, resident in a rural area (RURALSTATUS=1) or resident in a small metropolitan area (RURALSTATUS=2).

Marital: We construct a binary variable that takes on the value of 1 if the individual is married and the spouse is present, and 0 otherwise.

Race: We construct a categorical variable that takes on the value of 1 if the individual is coded as white, and 2 if non-white.

Hispanic: We construct a binary variable that takes on the value of 1 if the individual reports of Hispanic ethnicity, and 0 otherwise.

Veteran: We construct a binary variable that takes on the value of 1 if the individual is a veteran, and 0 otherwise.

3 APPENDIX C: Details for the Prediction Algorithms

In this section, we provide details of the algorithms we employed in the main text and we assess the robustness of the predictions to alternative definitions in determining "minimum wage workers". In addition, we provide suggestive evidence that misreporting of wages by workers does not have a major impact on our predictions by comparing models constructed using worker, and employer-reported wages using the January 1977 CPS Supplement.

Tree-based Machine Learning Tools

Random Forests

There is a multiplicity of ways to go beyond a single decision tree. The random forest is one of them (Breiman, 2001). It is a tree-based ensemble learning technique. It provides a way to overcome the bias-variance trade-off of the decision trees. In our case, it constructs a multitude of fully grown decision trees formed using different training bootstrap samples that predict the class of each observation. Using these predictions, it determines the class of the observation according to the majority vote. The final prediction of each observation, therefore, is the average of all the tree predictions.

Since each tree is fully grown, they are unbiased, yet the prediction variance is high. By averaging the predictions, the unbiasedness is preserved and the variance is diminished. To elaborate this point further, assume that the variance of the prediction of the decision trees is σ^2, and $\hat{f}^b(x)$ is the prediction of the decision tree that is formed using training sample b for given predictors, x. The random forest predicts the class of the observation by averaging the predictions $\hat{f}_{rf} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$, where B is the total number of trees. Then, the unbiasedness is retained, and if the predictions were independent from each other, the variance of the random forest predictions would be;

$$var(\hat{f_{rf}}) = var(\frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)) = (\frac{1}{B^2}) * B * \sigma^2 = \frac{\sigma^2}{B}. \quad (1)$$

However, the trees and the predictions are never uncorrelated. Especially, if one predictor has a very high predictive power, then the top node of all trees use it to split. Therefore, the variance

in equation 1 is, in fact, never achieved. To decrease the correlation of trees, instead of using all the predictors, we employ a randomly selected portion of the predictors at each split.[1] Using a fraction of the predictors might slow the learning process; though, with a large number of trees, it outperforms the random forest that uses all of the predictors at each step. In addition, thanks to the averaging, increasing the number of trees does not lead to overfitting; yet the prediction performance does not improve after a certain number of trees[2].

Gradient Tree Boosting

The boosting approaches the problem from a different angle. Instead of producing many fully grown trees, it starts by producing a relatively small tree. The initial tree is a weak learner, misclassifies many observations. The proceeding trees, also relatively small and weak learners, alter the data to predict the misclassified observations more accurately.

The gradient tree boosting developed by Friedman (2001) is one of the boosting techniques[3]. In our classification problem with only two classes, we choose the binomial log-likelihood loss function (or one half of the deviance). Hence, the loss function is;

$$L(y, p) = -(ylog(p) + (1 - y)log(1 - p)), \tag{2}$$

where y is the class of the observation ($y \in \{0, 1\}$). $p = P(Y = 1|X)$ where X indicates predictors. Then using the logistic transformation, we can write the loss function in terms of the model as:

$$L(y, F(X)) = -yF + log(1 + exp(F)). \tag{3}$$

Instead of fitting a single tree using 3, the boosting fits many weak trees sequentially. The main trick of the gradient tree boosting is that at step m, we replace the outcome with the negative gradient, given *m-1* boosting steps. To put it differently, the pseudo-response of the observation i,

[1] Using all predictors at every split is called "bagging".

[2] However, although increasing the number of trees does not lead to overfitting, individual trees themselves might overfit severely. As Segal (2004) shows, it is possible to improve the prediction by shrinking the tree size.

[3] For the purposes of this paper, we only describe the gradient tree boosting. Alternatively, one can also employ the AdaBoost algorithm. The gist of the AdaBoost algorithm is that at each step, observations are re-weighted so that misclassified (correctly classified) ones weigh slightly more (less) in the subsequent step. In our study, the AdaBoost performed slightly worse than the preferred model, so we omitted it. A very intuitive description of both the AdaBoost and gradient tree boosting algorithms can be found in Friedman et al. (2009).

$\tilde{y}_i$ is defined as:

$$\tilde{y}_i = -[\frac{\partial L}{\partial F}]_{F=F_{m-1}} = y_i - p_i. \tag{4}$$

and the tree at step m is fit to $\tilde{\boldsymbol{y}}$[4]. Then based on the fit, the new tree is added to the model according to the following formula;

$$F_m(X) = F_{m-1}(X) + \sum_{j=1}^{J} \gamma_{jm} \mathbf{1}(X \in R_{jm}), \tag{5}$$

where j is the terminal node of the tree, $F_{m-1}(X)$ is the model built at step $m-1$, and γ_{jm} is the optimal update coefficient that reduces the loss function the most at step m for the sub-space R_{jm}. Concretely, γ_{jm} takes higher values for subspaces that the mth tree fits relatively well. Note that due to the use of pseudo-responses, the minimization of loss function leads the mth tree to focus heavily on cases where F_{m-1} performs poorly. Only the combination of weak learners produce a strong learner, and most of the weak learners are meaningless by themselves.

Since we fit the tree to the negative gradient, the training error rate always decreases as the number of trees increases. Therefore, unlike the random forest, including many trees in the model can lead to overfitting in the gradient tree boosting. For regularization and better predictions, shrinkage techniques are employed. Instead of adding each new tree to the current model as in equation 5, every γ_{jm} is multiplied by a small positive number. The shrinkage parameter renders each learner even weaker and decelerates the learning process. Nevertheless, combined with sufficient number of trees, this can increase the model's predictive power significantly. Furthermore, using only a fraction of training set, known as sub-sampling, one can introduce another stochastic component and de-correlate trees; hence potentially decreasing the variance (Friedman, 2002).

Then, three main parameters that are user-specified in gradient boosting are: the size of each tree, the number of trees, and the shrinkage factor. We employ a 10-fold cross-validation procedure to determine each one of them simultaneously.

Before concluding this section, we note that although both random forests and gradient tree boosting build data-driven models that, in general, outperform a single tree, these models lack the interpretability. In the case of the random forest, due to the use of a fraction of variables at each

[4]For the least squares regressions, the negative gradient vector is the residuals at step m. Hence, the gradient boosting simply fits the regression tree to the current residuals at each step (James et al., 2013).

split, most of the individual trees cannot achieve the performance of a single tree, and the splits are not interpretable. On the other hand, the first tree of the gradient boosting can actually be the same as the decision tree; yet starting from the second tree, the interpretation of each tree is obscured due to the updating of the outcome variable or the loss function.

Elastic net

We also explore the performance of the logistic regression model. We use the elastic net regularization developed by Zou and Hastie (2005). The loss function of the logistic regression is:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) - ln(1 + exp(\beta_0 + \boldsymbol{x_i}\boldsymbol{\beta}))$$

The elastic net regularization adds the penalty term, $\lambda[(1-\alpha)||\beta||_2^2+\alpha||\beta||]$ to the loss function for regularization and decreasing the model complexity. $\lambda \geq 0$ and $\alpha \in [0,1]$ are tuning parameters picked by the cross-validation procedure to prevent overfitting. We purposefully build a very complex model and rely on the regularization to avoid overfitting. The model includes all the predictors, their two-way interactions, all the interactions with the quadratic age variable, and the cubic and quartic terms of the age variable.

Card and Krueger's linear probability model

We also apply the linear probability model analogous to the one employed by Card and Krueger (1995). While the former model has no polynomial or interaction terms, the right hand side variables of the latter model are a set of three-way interaction variables between teen, non-white, and gender indicators; three-way interaction variables between young adult (age 20-25), non-white, and gender indicators; three-way interactions of age, categorical education, and gender variables; quadratic and cubic terms of the age variable; indicator variables for Hispanic, and non-white individuals.

Robustness to alternative threshold values

In the main prediction model, we classified wage workers earning less than 125% of the minimum wage as relevant cases (minimum wage workers), and other wage workers as non-relevant cases. We use the threshold value primarily for description purposes, however it might be the case that

the arbitrarily selected threshold value has a qualitative effect on the predictions since it determines the value of the outcome variable. Taking the concern into account, we build prediction models for alternative threshold values. Since we use the predicted probabilities for sorting, we compare ranks of the predicted probabilities estimated by latter models with the main model using the 1979-2019 CPS-ORG. In addition, we also build a model to predict real wages, sort according to (the negative of) predicted wages, and compare the resulting order with the one obtained by the main prediction model.

In figure C.1 and table C.1, we show that selecting alternative threshold values produce virtually the same ordering. The rank correlation coefficients are always greater than 0.95. The coefficient is always greater than 0.99 for the threshold values (1.03, 1.1, 1.25, 1.5) used in Belman et al. (2015) when they try alternative threshold values in describing demographics and occupations of minimum wage workers. This implies that we obtain highly similar samples when we use alternative threshold values in the prediction model or predicted real wages for sorting. In other words, the high impact and the baseline groups would be essentially the same if we used another threshold wage level in defining the minimum wage workers.

Impact of misreporting error on the formation of groups

One issue that might affect the formation of predicted probability groups is the misreporting error. If the misreporting error of the hourly wage information is larger for certain groups, then the members of the groups might be incorrectly predicted to be more (or less) likely to be a minimum wage worker. An example that illustrates the issue is as follows: consider a case where the minimum wage is \$10, and two demographic groups have the same true wage distributions. The distributions are both normal with mean \$14.5 and standard deviation \$2. This implies that, in truth, 31.7% of both groups are minimum wage workers according to the convention (minimum wage workers are defined as those earning less than 125% of the minimum wage). Say, the first group reports their hourly wages with some error. The error would increase the observed standard deviation without affecting the mean. Then, the ML tools would predict the first group to be more likely to have minimum wage workers than the second group.[5]

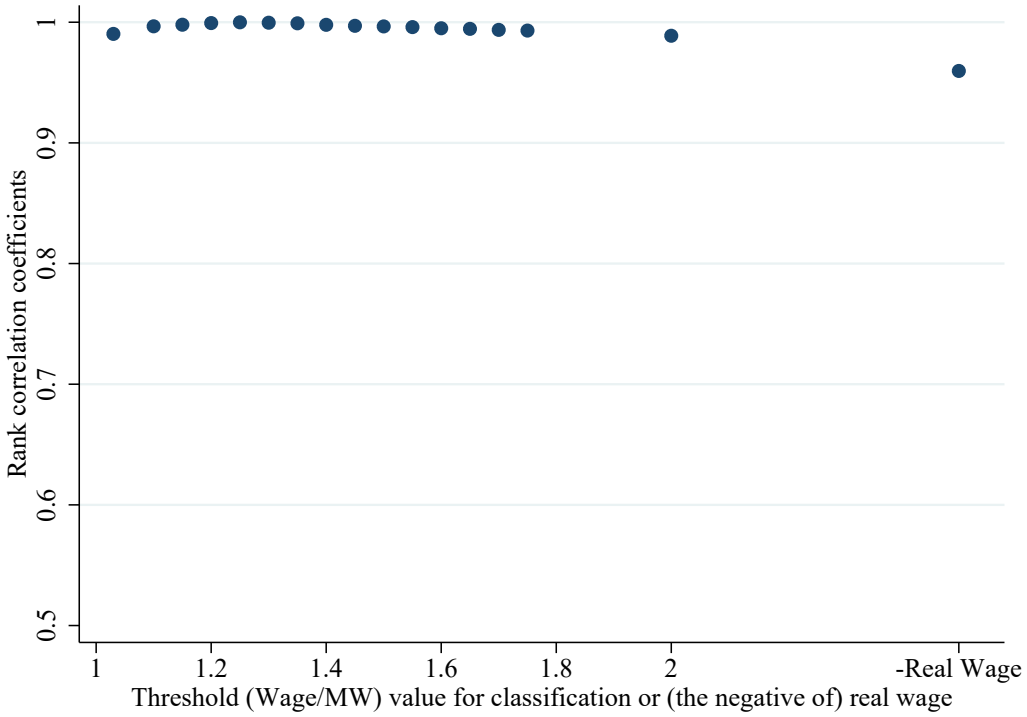
[5]The large rank correlation coefficient of the main predicted probabilities and the negative of predicted real wages in table C.1 imply that the severity of this kind of error is limited.

As suggestive evidence on the impact of misreporting error on the predictions, we employ the CPS 1977 January Supplement. In the data set, hourly wages of workers are asked to employers as well as to workers. We assess the severity of the misreporting error by comparing the prediction model constructed using worker-reported wages with that using employer-reported wages.

Specifically, we use 80% of the sample to train two models: the first model employs the worker-reported wage information, whereas the second model employs the employer-reported wage information. The predictors are the same as the model in the paper, except that the CPS 1977 does not have citizenship information, and the small metropolitan areas are defined as those with population less than 1,000,000 (rather than 500,000). After having constructed the models, we test the performances of each model using the remaining 20% of the original sample. Each model predicts the likelihood of being a minimum wage worker for each of the observations in the test sample. The similarity of the predictions suggests that the misreporting error negligibly affects the sorting.[6] The slope, the R^2, and the rank correlation coefficient are all very close to 1: they are 0.990, 0.989, 0.992, respectively. This suggests that using worker-reported wages results in similar predictions to using employer-reported wages.

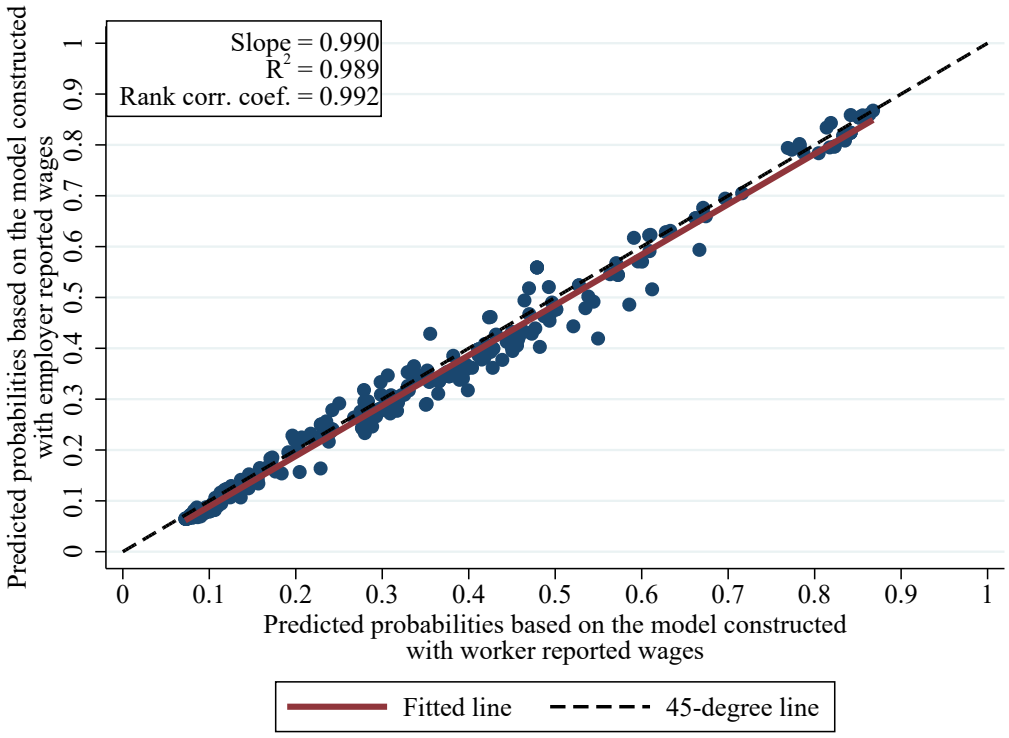
[6]One underlying assumption here is that the misreporting errors of employers and workers are independent from each other. More precisely, $w^{worker} = w + e^{worker}$, $w^{employer} = w + e^{employer}$ and e^{worker} is independent from $e^{employer}$.

Figure C.1: Rank Correlation Coefficients



Notes: The 1996-2017 CPS-ORG is employed. The graph shows the rank correlation coefficients of the predicted probabilities obtained from models constructed using alternative threshold $\frac{wage}{MW}$ values in classifying observations or from the model predicting the negative of real wage and the main predicted probabilities.

Figure C.2: Predictions of Models Based on Employer-Reported, and Worker-Reported Wages



Notes: In this figure, the models are fitted using the January 1977 CPS Supplement, separately for wages reported by workers versus employers. The graph plots the predicted probabilities of the model using employer-reported wages (y-axis) against those of the model using worker-reported wages (x-axis). Each marker indicates an observation, while the solid line shows the best linear fit. The dashed line shows the the 45 degree line.The R^2 of the fit, the slope of the line, and the rank correlation coefficients are reported in the top-left corner.

Table C.1: Rank correlation coefficients

Outcome variable:	Indicator for hourly wage below the threshold						-Real wage
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rank correlation coefficient	0.990	0.997	1.000	0.997	0.993	0.989	0.960
Threshold value	1.03	1.1	1.25	1.5	1.75	2	-

Notes: The 1996-2017 CPS-ORG is used. Estimated rank correlation coefficients of the main predicted probabilities and the predicted probabilities of the models that use alternative thresholds or the hourly real wage for the outcome variable. Columns (1)-(6) employ a binary outcome that takes on the value of 1 if the ratio of the real hourly wage to the minimum wage is below the specified threshold. Column (7) employs the negative of hourly real wage as the outcome, and predicts the worker's wage. Since the observations more likely to earn low wages are more likely to be minimum wage workers, we sort according to the negative of the real wage.

4 APPENDIX D: Importance of Participation in Flinn (2006)

The link between participation and welfare can be seen in a three state model of the labor market, such as Flinn (2006), where workers are either not participating in the labor market, participating and unemployed or participating and employed. We simplify Flinn (2006) by assuming firm homogeneity, and maintain the assumption that workers are identical except in their flow value of non-participation, ρV_i^n, which is distributed according to the cumulative distribution function F. Denote the lifetime value of being employed as V^e and unemployed as V^u, then when the minimum wage, m_w, binds we will have that

$$\begin{aligned}\rho V_i^n \sim & F \\ \rho V^u = & b + \lambda(m_w)(V^e - V^u) \\ \rho V^e = & m_w + \delta(V^u - V^e)\end{aligned}$$

where b is the replacement rate for unemployed workers, $\lambda(m_w)$ is the job finding rate which is allowed to vary with the minimum wage, and δ is the exogenous job destruction rate. Worker i participates if and only if $V^u >= V_i^n$, so the participation rate is given by $F(\rho V^u)$. If the value of non-participation is invariant to the minimum wage, then an increase in participation following a minimum wage increase is a sufficient condition for an in increase in the lifetime value of being unemployed, V^u , which factors in the joint impact of minimum wage changes on future wages conditional on employment, and on the probability of employment. This follows because the participation rate, $F(\rho V^u)$, increases if and only if V^u increases. Indeed if F is a twice differentiable continuous function and $sup(\rho V_i^n) > \rho V^u$, i.e. participation is not 100%, then an increase (decrease) in participation is both a necessary and sufficient condition for an increase in the lifetime value of being unemployed. However, in a more general setting where some groups do have close to 100% participation rates then aggregate participation changes are more informative about welfare changes for marginal rather than infra-marginal groups of workers.

5 APPENDIX E: Stacked Regression Analysis

This section provides an alternative to our baseline panel specification building on the stacked regression approach proposed by Cengiz et al. (2019) (see their Appendix D). We begin with each of the 172 prominent minimum wage events, and create an event-specific dataset that includes the treated state, and all other clean control states for an 8-year panel by event time ($t = -3, ..., 4$), with the minimum wage increase occurring at $t = 0$. Clean controls are those states without any non-trivial minimum wage increase within the 8-year event window. Some of these events have other minimum wage increases that happened prior to date $t = 0$, i.e., during $t \in \{-3, -1\}$. To avoid contamination from past increases, we additionally exclude all events that have another minimum increase in the three years prior. This produces a set of 47 clean treatment events.[7]

Then we stack all of the event-specific data to calculate an average effect across all the events using the a single set of treatment effects β_τ:

$$Y_{hst} = \sum_{\tau=-3}^{4} \beta_\tau treat_{hst}^{\tau} + \Omega_{hst} + \mu_{hs} + \rho_{ht} + u_{hst}$$

we calculate the standard errors by clustering at the state-level. Note that all the state and time fixed effects (μ_{hs}, ρ_{ht}) are fully interacted by event (h), which means we are only using within-event (and not between-event) variation for identification.[8] Here Ω_{hst} is a vector of (event-specific) controls for federal minimum wage events, as well as non-prominent (\$0.25 or less, or affecting less than 2% of workers) state-level events.

By aligning events by event-time (and not calendar time), using only "and by using only within-event variation, it is equivalent to a setting where the events happen all at once, and are not staggered. Moreover, by dropping all states with any events within the 8 year event window from the control set, we guard against bias due to heterogeneous treatment effects. Together, these two features prevent negative weighting of some events that may occur with a staggered design

[7] Cengiz et al. (2019) did not exclude any events, but controlled separately for other events within the window. Using that approach produces similar point estimates. However, unlike in Cengiz et al. (2019), here the estimates without exclusion show pre-existing trends in minimum wages, suggesting contamination (this likely stems from the inclusion of more recent data with longer phase-in periods). Excluding events with other minimum wage increases in the pre-period directly removes this problem.

[8] In Cengiz et al. (2019), the reported standard errors are clustered at the state-event level, even though estimated standard errors were very similar using state-level clustering. The state-event level clustering assumes independence of errors across event-specific sub-datasets. Since the same states appear in many sub-datasets it is more appropriate to cluster the standard errors by state, which we apply here. We thank Kirill Borusyak for pointing this out.

(Sun and Abraham (2020)). This is shown formally in Gardner (2021), who derives the implicit weight placed on each event, h, in our "stacked regression" model (see Appendix A of that paper). The weight for event h depends on attributes that determine the variance in treatment (number of control units used in event h, and the share of observations in dataset h as a share of the overall stacked dataset), as expected. Importantly, these weights do not suffer from negative weighting.

Moving to the stacked-by-event approach (column 2 and 4 in Table E.1) continues to produce a sizeable and statistically significant positive wage effect; the somewhat larger point estimate reflects the presence of additional minimum wage increases in the post-treatment period. The employment, unemployment and participation effects that are statistically indistinguishable from zero, like in our baseline model. The implied minimum wage elasticities and own wage elasticities are also small, and indistinguishable from zero.

Table E.1: Impact of the Minimum Wage: Stacked Data Estimates

	(1)	(2)	(3)	(4)
Δ wage (%)	0.021***	0.034***	0.015***	0.022***
	(0.003)	(0.006)	(0.003)	(0.005)
Δ employment (% pt)	0.001	-0.004	0.001	0.001
	(0.002)	(0.004)	(0.001)	(0.003)
Δ unemployment (% pt)	-0.001	-0.001	-0.000	0.000
	(0.001)	(0.002)	(0.000)	(0.001)
Δ participation (% pt)	0.000	-0.005	0.001	0.001
	(0.002)	(0.004)	(0.001)	(0.003)
Employment Elas. w.r.t Min. Wage	0.035	-0.079	0.031	0.014
	(0.075)	(0.087)	(0.033)	(0.052)
Employment Elas. w.r.t Wage	0.159	-0.315	0.192	0.086
	(0.342)	(0.337)	(0.195)	(0.316)
Number of events	172	47	172	47
Number of observations	7,854	40,284	7,854	40,284
Number of individuals in the sample	7,581,999	34,433,189	23,011,491	105,028,891
Mean employment	0.347	0.342	0.431	0.431
Mean unemployment	0.057	0.054	0.045	0.045
Mean participation	0.404	0.396	0.476	0.475
Group	High Prob.	High Prob.	High Recall	High Recall
Estimation Method	Baseline	Pooled-Stacked	Baseline	Pooled-Stacked

Notes: The table reports the effects of the minimum wage on labor market outcomes based on the event study analysis using state-level minimum wage changes between 1979 and 2019. The table reports five year averaged post-treatment estimates for each key labor market outcome: percent change in wages and the change in employment to population, unemployment to population, and labor force participation rate. We also report the employment elasticity with respect to the minimum wage and the employment elasticity with respect to the wage, which is the ratio of the percent change in employment and wage. To calculate the percent change in employment we divide the change in employment to population by the mean employment to population rate preceding the minimum wage hikes (reported at the bottom of the table). The line on the number of observations shows the number of quarter-state cells used for estimation in the baseline estimation method or the event-state cells used in the pooled-stacked and manually averaged methods, while the number of individuals refers to the underlying CPS sample used to calculate labor market outcomes in these cells. Columns (1) and (2) show estimates for the high-probability group, which captures 10% of the population with highest predicted probability. Columns (3) and (4) show estimates for the high-recall group, which consists of individuals whose predicted probability is above 11% - a threshold which leads to a 75% recall rate of minimum wage workers. Columns (1) and (3) use the baseline panel method, using 172 state-level minimum wage changes. Columns (2) and (4) use the stacked event study approach, using 47 events with no minimum wage increases in the 3 years prior to the event. All regressions are weighted by state-quarter population. Robust standard errors in parentheses are clustered by state; significance levels are * 0.10, ** 0.05, *** 0.01.

References

Belman, Dale, Paul Wolfson, and Kritkorn Nawakitphaitoon. 2015. "Who Is Affected by the Minimum Wage?" *Industrial Relations: A Journal of Economy and Society*, 54(4): 582–621.

Breiman, Leo. 2001. "Random forests," *Machine learning*, 45(1): 5–32.

Card, David and Alan B. Krueger. 1995. *Myth and Measurement: The New Economics of the Minimum Wage*, New Jersey: Princeton University Press.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs*," *The Quarterly Journal of Economics*, 134(3): 1405–1454.

Feenberg, Daniel and Jean Roth. 2007. "CPS labor extracts 1979–2006."

Flinn, Christopher J. 2006. "Minimum Wage Effects on Labor Market Outcomes under Search, Matching, and Endogenous Contact Rates," *Econometrica*, 74(4): 1013–1062.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. 2020. "Current Population Survey: Version 7.0 of Dataset," *Integrated Public Use Microdata Series (IPUMS): Minneapolis*, DOI: http://dx.doi.org/https://doi.org/10.18128/D030.V7.0.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine," *Annals of statistics*: 1189–1232.

——— 2002. "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, 38(4): 367–378.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: New York, NY: Springer-Verlag New York.

Gardner, John. 2021. "Two Stage DiD and Taming the DiD Revolution," *Working paper*.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*, 6: Springer.

Segal, Mark R. 2004. "Machine learning benchmarks and random forest regression," *Center for Bioinformatics & Molecular Biostatistics*.

Sun, Liyang and Sarah Abraham. 2020. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*.

Vaghul, Kavya and Ben Zipperer. 2016. "Historical state and sub-state minimum wage data," *Washington Center for Equitable Growth Working Paper*.

Zou, Hui and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320.