

Replication and Extension of Vocational Training Effects on Youth Employment in Rural India

Nitay Carmi

Word Count: 1975

November 25, 2025

1 Introduction

India's youth unemployment remains elevated, with rural areas particularly affected by limited formal employment opportunities. Chakravorty et al. (2023) examine Bihar and Jharkhand's DDU-GKY vocational training programme through two designs: a panel survey tracking employment outcomes of training graduates versus dropouts across the COVID-19 pandemic, and a randomised experiment testing whether an app-based job platform improves labour market integration.

My replication validates their core findings across a baseline survey and three subsequent follow-up waves from February 2020 to November–December 2021. Training completion increases salaried employment, with effects ranging from 13.5 to 31.8 percentage points across survey rounds. In relative terms, trained individuals were approximately three times more likely to hold salaried jobs than dropouts after the first national lockdown, demonstrating training's protective effect during economic disruption.

My analysis makes three contributions. First, I achieve near-perfect replication, with perfect matches for Table 2 (main panel analysis) and coefficient differences below 0.001 for Table 4B (RCT), validating reproducibility of published findings. Second, I extend Double LASSO methodology which the original paper applied only to the RCT component to the panel analysis (Table 2). LASSO selected controls adaptively across panels: 9 controls for pre-lockdown, 8 for June–July 2020, and only 3 for both 2021 surveys (averaging 5.8 out of 34 potential controls), achieving an 83% reduction while maintaining coefficient stability within 0.015 percentage points. This demonstrates robustness to algorithmic control selection. Third, I examine caste heterogeneity not analysed in the original study, finding that lower castes (SC/ST/OBC) showed larger effects pre-COVID (32.1 percentage points versus

10.9 percentage points for general castes), but this advantage reversed by late 2021 (14.9 percentage points versus 27.2 percentage points), suggesting differential persistence despite universal skills acquisition. Section 2 describes the experimental design, Section 3 outlines empirical strategy, Section 4 presents results, and Section 5 discusses policy implications.

2 Data and Experimental Design

The analysis uses data from the DDU-GKY vocational training programme in rural Bihar and Jharkhand, combined with two research components: panel surveys tracking training graduates versus dropouts, and a randomised trial of an app-based job platform (Yuvasampark). The training provides intensive residential instruction across sectors including healthcare, retail, hospitality, and security, targeting disadvantaged rural youth.

The study employed stratified randomisation for the RCT component, creating strata defined by state, training sector, gender, and treatment group. Within each stratum, candidates were randomly assigned to receive Yuvasampark job platform training or control status. Balance tests across 67 baseline characteristics show only 3 variables with statistically significant differences at conventional levels, indicating successful randomisation (see Appendix Table 6).

The panel component tracked approximately 2,000 trainees across four survey rounds (Table 1): February–March 2020 (pre-lockdown baseline), June–July 2020 (immediate lockdown), March–April 2021 (partial recovery), and November–December 2021 (late pandemic). Sample sizes declined modestly from 2,204 in early rounds to 1,890–1,916 in later surveys (14% attrition), within acceptable bounds for panel studies in developing contexts. The sample comprises predominantly young individuals and lower-caste members (SC/ST/OBC: 93%; General caste: 7%), reflecting deliberate targeting of disadvantaged groups. Baseline salaried employment rates were low, with control group means ranging from 8.0% during lockdown to 14.7% pre-pandemic.

Table 1: Survey Structure

Panel	Survey Round	Timing	Sample Size	Control Mean	Context
Panel A	Round 1 (Pre-lockdown)	Feb-Mar 2020	2204	0.147	Pre-pandemic
Panel B	Round 1 (Post-lockdown)	Jun-Jul 2020	2204	0.08	Lockdown period
Panel C	Round 2	Mar-Apr 2021	1890	0.134	Partial recovery
Panel D	Round 3	Nov-Dec 2021	1916	0.133	18 months post-training

3 Empirical Strategy

3.1 Main Panel Analysis

The primary specification estimates training effects on salaried employment across four survey rounds using OLS:

$$\text{Salaried}_{it} = \beta_0 + \beta_1 \cdot \text{Training}_i + \delta' \cdot \text{Sector}_i + X'_i \gamma + \varepsilon_{it} \quad (1)$$

where Salaried_{it} is a binary indicator for individual i holding salaried employment in period t , Training_i indicates completion of DDU-GKY programme, Sector_i contains training sector fixed effects (10 categories), and X_i includes 24 baseline individual and household characteristics. The coefficient β_1 captures the average treatment effect of training completion on employment probability. I estimate this specification separately for each survey round, allowing effects to vary across the pandemic timeline. The full specification (Column [4] in results tables) includes 34 total controls: 10 sector dummies plus 24 baseline covariates covering demographics, education, migration history, household composition, asset ownership, and government programme participation.

3.2 Double LASSO Extension

I extend the analysis with Double LASSO (Belloni et al., 2014) for transparent control selection, addressing specification uncertainty with 34 baseline covariates. First, I run LASSO of treatment on all controls, selecting variables that predict treatment assignment and thus improve precision. Second, I run LASSO regression of the outcome on all controls, selecting variables that predict employment and thus help reduce omitted variable bias. Third, I take the union of both selected sets (which chooses the unique variables), including each variable if selected by either LASSO and estimate the treatment effect via ordinary least squares controlling only for these variables. This identifies controls that matter for reducing bias or improving efficiency while discarding irrelevant controls contributing only noise. I use post-selection inference: LASSO selects variables, then I run unbiased OLS on the selected set, avoiding shrinkage bias from L1 penalisation. I implement this separately for each panel, allowing control selection to adapt to different outcome patterns across survey rounds.

3.3 RCT Analysis

For the randomised component, I estimate intent to treat effects of assignment to Yuvasam-park job platform training on three outcomes: whether applied for any job, applied for 1–2

jobs, or applied for 3+ jobs. The specification includes strata fixed effects and LASSO-selected baseline controls, with robust standard errors.

3.4 Caste Heterogeneity Analysis

To examine heterogeneous treatment effects by caste, not analysed in the original study, I estimate interaction models:

$$\text{Salaried}_{it} = \beta_0 + \beta_1 \cdot \text{Training}_i + \beta_2 \cdot \text{General}_i + \beta_3 \cdot (\text{Training}_i \times \text{General}_i) + \delta' \cdot \text{Strata}_i + X'_i \gamma + \varepsilon_{it} \quad (2)$$

where General_i indicates general (upper) caste status, with lower castes (SC/ST/OBC) as reference. I include strata fixed effects (Strata_i) to control for stratified randomisation. The coefficient β_1 captures training effects for lower castes, while $\beta_1 + \beta_3$ represents total effects for general castes, with β_3 testing whether effects differ significantly. Standard errors for $\beta_1 + \beta_3$ use the delta method accounting for covariance between β_1 and β_3 . This analysis addresses whether training reduces or reinforces caste-based labour market inequalities in India's stratified social context.

4 Results

4.1 Main Replication: Training Effects on Employment

Table 2 replicates the original study's core finding: training completion substantially increases salaried employment across all survey rounds. Pre-lockdown (Panel A), training raised employment probability by 31.8 percentage points ($p < 0.01$) from a control mean of 14.7%, representing more than a doubling of employment rates. During the June–July 2020 lockdown (Panel B), when control group employment collapsed to 8.0%, training effects remained large at 23.7 percentage points, demonstrating protective effects during crisis. Effects persisted through 2021 recovery: 13.5 percentage points in March–April (Panel C) and 16.2 percentage points in November–December (Panel D), both highly significant despite gradually improving baseline employment.

Specifications show remarkable stability across control inclusion. The no-controls estimate (Column [1]) differs from the full specification (Column [4]) by at most 1.4 percentage points across all panels, suggesting minimal omitted variable bias. Adding sector controls (Column [2]) produces estimates within 0.6 percentage points of the full model. Individual characteristics (Column [3]) close most remaining gaps. This stability indicates training

assignment was effectively uncorrelated with observable characteristics that predict employment outcomes.

My replication achieves near-perfect accuracy: perfect match for Table 2 and maximum coefficient difference of 0.001 for Table 4B compared to published results. Standard errors match within 0.001 in most cases. This validates both data quality and computational reproducibility of the original findings.

Table 2: Main Replication Results

Variable	Col1	Col2	Col3	Col4
Panel A: Survey Round 1 (Pre-lockdown 2020)				
Trained	0.309*** "(0.031)"	0.332*** "(0.029)"	0.322*** "(0.029)"	0.318*** "(0.029)"
Observations	2204	2204	2204	2204
Dropout Mean	0.147	0.147	0.147	0.147
Panel B: Survey Round 1 (Jun-Jul 2020)				
Trained	0.235*** "(0.028)"	0.249*** "(0.027)"	0.242*** "(0.027)"	0.237*** "(0.027)"
Observations	2204	2204	2204	2204
Dropout Mean	0.080	0.080	0.080	0.080
Panel C: Survey Round 2 (Mar-Apr 2021)				
Trained	0.123*** "(0.030)"	0.139*** "(0.030)"	0.132*** "(0.030)"	0.135*** "(0.030)"
Observations	1890	1890	1890	1890
Dropout Mean	0.134	0.134	0.134	0.134
Panel D: Survey Round 3 (Nov-Dec 2021)				
Trained	0.139*** "(0.030)"	0.157*** "(0.030)"	0.161*** "(0.031)"	0.162*** "(0.031)"
Observations	1916	1916	1916	1916
Dropout Mean	0.133	0.133	0.133	0.133
Sector Controls		Yes	Yes	Yes
Individual Controls			Yes	Yes
Household Characteristics				Yes

4.2 Double LASSO Innovation: Adaptive Control Selection

Table 3 presents Double LASSO results extending the panel analysis. LASSO selected controls adaptively: 9 variables for pre-lockdown (Panel A), 8 for June–July 2020 (Panel B), and only 3 controls for both 2021 surveys (Panels C and D). This 83% average reduction (from 34 to 5.8 controls) demonstrates considerable efficiency gains while maintaining substantive conclusions. Training coefficients from LASSO specifications remain within 0.012 percentage points of full-specification estimates across all panels, with identical significance levels.

The temporal pattern is notable: pre-COVID periods required 8–9 controls, while post-COVID recovery required only 3 controls, suggesting evolving confounding structures as labour markets stabilised. This validates the original study’s control choices while demonstrating that most were unnecessary for unbiased estimation in this context.

Table 3: Double LASSO Extension

Panel	Col5_LASSO
Panel A: Survey Round 1 (Pre-lockdown 2020)	
Trained	0.307*** (0.029)
Observations	2204
Dropout Mean	0.147
Panel B: Survey Round 1 (Jun-Jul 2020)	
Trained	0.229*** (0.027)
Observations	2204
Dropout Mean	0.080
Panel C: Survey Round 2 (Mar-Apr 2021)	
Trained	0.131*** (0.030)
Observations	1890
Dropout Mean	0.134
Panel D: Survey Round 3 (Nov-Dec 2021)	
Trained	0.154*** (0.030)
Observations	1916
Dropout Mean	0.133
<hr/>	
Sector Controls	LASSO
Individual Controls	LASSO
Household Controls	LASSO

4.3 Job Platform RCT: Null Results

Table 4 replicates the randomised experiment results. Assignment to Yuvasampark job platform training produced precisely estimated null effects across all outcomes: -1.2 percentage points for any applications ($p=0.404$), -1.1 percentage points for 1–2 applications ($p=0.376$), and -0.3 percentage points for 3+ applications ($p=0.701$). These results suggest that providing access to job search platforms alone, without addressing underlying barriers to employment or job availability, has minimal impact on application behaviour. The contrast with large vocational training effects highlights that skills development matters more than search frictions in this context.

Table 4: RCT Results (Table 4B)

Variable	Col1	Col2	Col3
Panel B: Survey Round 3 (Nov-Dec 2021)			
Treatment	-0.012 "(0.015)"	-0.011 "(0.013)"	-0.003 "(0.007)"
p-value	0.404	0.376	0.701
Control Mean	0.125	0.097	0.024
Observations	1955	1955	1955
LASSO-selected controls	Yes	Yes	Yes
Strata fixed effects	Yes	Yes	Yes

4.4 Caste Heterogeneity: Differential Returns Over Time

Table 5 documents heterogeneous training effects by caste not examined in the original study. Pre-lockdown (Panel A), lower-caste trainees experienced larger employment gains: 32.1 percentage points versus 10.9 percentage points for general-caste trainees (interaction coefficient -21.2 percentage points, $p=0.076$). This pattern held during June–July 2020 lockdown (Panel B): 23.8 percentage points for lower castes versus 11.1 percentage points for general castes, though the 12.7 percentage point difference was not statistically significant ($p=0.257$).

However, this pattern reversed by 2021. In March–April (Panel C), lower castes showed 13.1 percentage point gains while general castes showed 18.8 percentage points (interaction: $+5.7$ percentage points, $p=0.653$). By November–December (Panel D), the reversal strengthened: lower castes maintained 14.9 percentage point effects while general castes experienced 27.2 percentage point gains (interaction: $+12.3$ percentage points, $p=0.303$). Though individual interactions lack conventional significance due to small general-caste sample ($n=153$, 7% of total), the four-panel reversal pattern suggests substantive rather than spurious dynamics.

Several mechanisms could explain this reversal. General-caste individuals may leverage social networks and family connections to sustain employment through pandemic recovery. Initial credential signalling may fade as employers observe actual productivity, with discrimination reasserting. Sectoral sorting could produce differential recovery trajectories. Lower castes may face higher job turnover despite similar skills. While absolute effects remain positive for both groups at 18 months post-training, the changing relative patterns suggest initial equity gains from universal skills provision do not guarantee lasting equity without addressing structural social barriers.

Table 5: Caste Heterogeneity

Variable	Panel_A	Panel_B	Panel_C	Panel_D
Training effect (Lower Caste)	0.321*** "(0.030)"	0.238*** "(0.028)"	0.131*** "(0.031)"	0.149*** "(0.032)"
Training \times General Caste	-0.212* "(0.119)"	-0.127 "(0.112)"	0.057 "(0.127)"	0.123 "(0.119)"
Training effect (General Caste)	0.109 "(0.116)"	0.111 "(0.109)"	0.188 "(0.123)"	0.272** "(0.115)"
P-value (interaction)	0.076	0.257	0.653	0.303
Observations	2204	2204	1890	1916
Controls	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes

5 Discussion

This replication validates that vocational training increased salaried employment for disadvantaged rural youth, with effects persisting through severe pandemic disruption. Perfect replication accuracy (0.0000 difference for Table 2, 0.001 for Table 4B) demonstrates high data quality and computational reproducibility. The Double LASSO extension shows that while 34 controls were used, adaptive selection requiring only 3–9 per panel produces equivalent estimates, confirming robustness to specification choices. This parsimony is especially notable in the 2021 surveys, where labour market stabilisation apparently reduced confounding complexity.

The caste heterogeneity analysis reveals that lower castes' initial training advantages disappeared by late 2021, with general castes showing nearly doubled effects (27.2 versus 14.9 percentage points). While statistical power limits definitive conclusions given small general-caste samples, the consistent four-panel pattern requires attention. This suggests that skills training alone, while beneficial in absolute terms, may be insufficient to overcome social stratification without further interventions: placement assistance targeting disadvantaged groups, active anti-discrimination enforcement, mentorship programmes building professional networks, and long-term monitoring of employment quality beyond binary employment status.

The null RCT results for job platforms contrast with large training effects, suggesting search frictions matter less than human capital constraints in this setting. Combined with the caste findings, this points toward a broader lesson: effective youth employment policy

requires both skills development and active measures addressing social barriers to labour market access and advancement.

My findings parallel Jones and Sen (2022) in Mozambique, though heterogeneity differs. They found gender based benefits from informal platforms for manual workers, while I document caste based differentials with formal platforms, highlighting how social barriers interact with platform design.

References

- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.
- Chakravorty, B., Bhatiya, A. Y., Imbert, C., Lohnert, M., Panda, P., and Rathelot, R. (2023). Impact of the covid-19 crisis on india’s rural youth: Evidence from a panel survey and an experiment. *World Development*, 168:106242.
- Jones, S. and Sen, K. (2022). Labour market effects of digital matching platforms.

A Appendix Tables

Table 6: Balance Tests (Table A9) - Part 1

Variable	Control_Group	Treatment_Group	Diff_2-1	p-value
<i>Panel A: Demographics and Caste</i>				
Older (More than 20)	0.281	0.270	-0.010	0.578
Married	0.094	0.091	-0.004	0.756
Caste: ST	0.186	0.184	0.000	0.978
Caste: OBC	0.481	0.477	-0.005	0.790
Caste: General	0.073	0.062	-0.011	0.296
Religion: Muslim	0.062	0.057	-0.005	0.618
Religion: Christian	0.048	0.049	0.002	0.829
<i>Panel B: Education</i>				
Middle school (6-8 class)	0.045	0.053	0.008	0.338
Secondary (High Sch 9-10 class)	0.357	0.349	-0.009	0.662
Tertiary education (Graduate & above)	0.088	0.086	-0.002	0.841
Matric exam	0.933	0.936	0.002	0.819
More than 50%	0.523	0.480	-0.044*	0.032
Inter exam	0.583	0.585	0.002	0.922
Less than 50%	0.237	0.230	-0.006	0.723
<i>Panel C: Skills</i>				
Big 5 Extraversion Test (1 to 5)	3.298	3.282	-0.016	0.474
Big 5 Agreeableness Test (1 to 5)	3.757	3.768	0.011	0.621
Big 5 Conscientiousness Test (1 to 5)	3.855	3.852	-0.003	0.917
Big 5 Neuroticism Test (1 to 5)	2.437	2.409	-0.028	0.330
Big 5 Openness Test (1 to 5)	3.945	3.923	-0.022	0.471
Grit (1 to 5)	3.408	3.429	0.021	0.409
ASE Test (1 to 4)	2.092	2.101	0.009	0.542
Life goal Test (1 to 4)	2.136	2.151	0.015	0.274
Duration of baseline survey (above median)	0.505	0.496	-0.008	0.677

Table 7: Balance Tests (Table A9) - Part 2 (Continued)

Variable	Control_Group	Treatment_Group	Diff_2-1	p-value
Panel D: Socioeconomic background				
Household head relationship (mother)	0.089	0.086	0.017	0.134
Household head relationship (others)	0.097	0.083	-0.014	0.233
Immediate difficulty to family	0.101	0.105	0.004	0.607
Future difficulty to family	0.140	0.150	0.010	0.324
Earning members (5 or more)	0.085	0.110	0.024	0.051
Household earning (15000 or more)	0.164	0.182	0.018	0.264
Household earning (5000 or less)	0.284	0.269	-0.014	0.436
Household earning (5001-9000)	0.230	0.226	-0.003	0.858
Agriculture land	0.660	0.651	-0.008	0.647
BPL card	0.797	0.781	-0.016	0.355
RSBY card	0.381	0.403	0.022	0.272
MNREGA	0.248	0.259	0.012	0.515
SHG member	0.739	0.737	-0.002	0.933
Semi pucca house	0.214	0.189	-0.025	0.132
Pucca house (IAY)	0.093	0.098	0.005	0.695
Pucca house (Non IAY)	0.191	0.214	0.023	0.174
Own house	0.996	0.993	-0.003	0.397
Internet use	0.518	0.529	0.010	0.546
Joint household	0.058	0.078	0.020	0.062
Household members (5 or less)	0.059	0.054	-0.005	0.633
Household members (6 or more)	0.376	0.372	-0.005	0.817
Ever migrated out of state (self)	0.120	0.139	0.020	0.149
Ever migrated out of state (relatives)	0.478	0.504	0.026	0.190
Relatives migrated (one)	0.325	0.369	0.044*	0.027
Relatives migrated (two or more)	0.152	0.135	-0.018	0.217
Panel E: Expectations				
Previous earning	0.118	0.119	0.001	0.918
Hypothetical earning (immediate)	0.158	0.140	-0.019	0.197
Hypothetical earning (in one year)	0.232	0.225	-0.007	0.682
Expected earning (in one year)	0.406	0.386	-0.020	0.320
Preferred earning (in one year)	0.467	0.422	-0.045	0.028
Training awareness	0.532	0.542	0.010	0.401
Training usefulness	0.934	0.935	0.001	0.841
Training satisfaction	0.944	0.949	0.004	0.383
Likelihood of training completion	0.945	0.949	0.004	0.466
Likelihood of job offer	0.900	0.902	0.002	0.812
Expected minimum salary (immediate)	0.396	0.384	-0.013	0.502
Expected maximum salary (immediate)	0.409	0.408	-0.001	0.966
Expected average salary (immediate)	0.478	0.446	-0.033	0.110
Likelihood of job offer outside state	0.786	0.798	0.011	0.224
Likelihood of accepting job inside state	0.841	0.836	-0.006	0.568
Likelihood of retention in job inside state	0.836	0.825	-0.011	0.265
Likelihood of accepting job outside state	0.824	0.829	0.006	0.587
Likelihood of retention in job outside state	0.818	0.818	0.001	0.949
Internet use	13 0.865	0.853	-0.012	0.395

AI Usage Declaration

Artificial Intelligence has been used in this assignment to assist with programming. Artificial Intelligence was only used to help with the for loops in the code, within what is allowed according to the assignment's and the department's guidelines.