Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Big-data pilot zones and provincial education and registered unemployment in China, 2010–2023

Qingxin Wang
Anhui Wenda University of Information Engineering
Jingdong Luan
Anhui Agricultural University
Rui Jin
Anhui Wenda University of Information Engineering
Hao Huang
HaoHuang@alicloude.cn

Chinese Academy of Social Sciences

Research Article

Keywords: big-data pilot zones, educational attainment, registered unemployment, digital economy, difference-in-differences, China

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11766037/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

China's National Big Data Comprehensive Pilot Zones were established in 2016 against a rapid expansion of higher education. We examine whether pilot-zone designation was associated with provincial educational attainment and registered urban unemployment, using official annual statistics for 31 mainland provincial-level units. The education panel covers 2010–2023; the unemployment-rate panel ends in 2021. Two-way fixed-effects estimates compare an inclusive coding of all ten host provinces with a stricter specification that excludes Guangdong and Liaoning, where the zones cover only part of the province. For 2010–2021, the estimated association with the registered unemployment rate is 0.382 percentage points (95% confidence interval [CI] 0.083 to 0.682) under the inclusive mapping and 0.317 points (95% CI 0.013 to 0.620) under the strict mapping. Estimates for 2010–2019 are smaller and less precise. Education estimates are sensitive to province-specific trends, and both conventional and interaction-weighted event studies reject the pre-treatment restrictions for education. An exploratory sum of the within-year standardised education and unemployment measures is mainly associated with the unemployment component and is sensitive to specification. These aggregate data cannot measure worker–job educational mismatch or establish the causal effect of pilot designation; the results instead show how the provincial indicators differed across zones and comparison areas and how strongly the estimates depend on measurement and design choices.

Introduction

China's higher-education expansion has increased the number of graduates seeking work in a labour market that is also being reshaped by digital technologies. Overeducation occurs when a worker's schooling exceeds the schooling required by the job; it is associated with wage penalties, although educational mismatch and skill mismatch are not identical (Duncan and Hoffman, 1981; Allen and van der Velden, 2001; McGuinness, 2006). Concerns about the capacity of the Chinese labour market to absorb graduates accompanied the transition to mass higher education (Bai, 2006; Li et al., 2014). Figure 1 places the expansion in a longer national context.

Digital technologies could change both the demand for skills and the way workers find jobs. The automation of routine tasks may displace some occupations, while new technologies can complement non-routine work and create new tasks (Autor et al., 2003; Autor, 2015; Acemoglu and Restrepo, 2020). Internet-based search may also give workers access to vacancies beyond their immediate networks (Autor, 2001; Kuhn and Mansour, 2014). None of these channels implies a fixed direction for the number of well-matched jobs: the outcome depends on how skills, vacancies, worker mobility and institutions change together.

China's National Big Data Comprehensive Pilot Zones offer a setting in which to examine aggregate labour-market patterns around a place-based digital policy. A study of the same programme considers firm employment (Li and Li, 2025), and another reports a positive association with firms' labour-income share (Xu et al., 2024). Cao and Cao (2024) find that individual digital skills and regional digital

development are associated with less education–job mismatch in Chinese survey data. Research on the separate smart-city pilot programme reports lower employment pressure (Ling et al., 2023). These studies use different interventions, units and outcomes; provincial education and registered unemployment provide a complementary view but do not link a worker's qualification to the requirements of a job.

We compare pilot and non-pilot provinces before and after designation, using annual education and registered-unemployment series from the China Statistical Yearbook. Because the Pearl River Delta and Shenyang zones do not cover their entire host provinces, we present estimates both including Guangdong and Liaoning as treated provinces and excluding them. We also examine changes in estimates when the sample period is restricted and province-specific trends are included. These comparisons make the dependence of the results on geography, time coverage and pre-existing trajectories visible.

In the full unemployment-rate sample, pilot designation is associated with higher registered unemployment relative to comparison provinces, but the pre-2020 estimates are smaller and their confidence intervals include zero. Education estimates are likewise imprecise and sensitive to trends, and pre-treatment education trajectories differ. The study therefore informs the interpretation of provincial indicators around digital-policy designation rather than providing a direct test of individual overeducation or a definitive causal estimate.

Digitalisation and provincial labour-market indicators

Education, jobs and digital change

Task-based theories emphasise that technology can substitute for some activities and complement others (Spitz-Oener, 2006; Acemoglu and Autor, 2011). The effects on a worker depend on the tasks in the job and on the availability of new work, not simply on the worker's schooling. Shifts in the occupational structure may be accompanied by changes in the relative demand for social and technical skills (Goos et al., 2014; Deming, 2017). Online job search can lower information costs, but the distribution of any gains depends on who has access to vacancies and the skills employers seek. Digital infrastructure is consequently one element of a broader institutional setting (World Bank Group, 2016).

The National Big Data Comprehensive Pilot Zones combined initiatives in infrastructure, data governance and industrial development. A change in the share of residents with college education could reflect schooling, migration or demographic composition. A change in registered unemployment could reflect labour-market conditions or the propensity and eligibility to register. Neither series alone identifies the availability of jobs that use graduates' qualifications. Other Chinese digital-policy studies consider different interventions or populations, including rural occupational transformation (Xiong and Sui, 2025); differences in outcomes and units of analysis matter when comparing findings.

Interpreting aggregate measures

A direct measure of overeducation pairs a worker's education with the educational requirements of that worker's job (Chevalier, 2003; Leuven and Oosterbeek, 2011). The provincial series used here contain neither this pairing nor a measure of the educational composition of vacancies. A higher college-educated population share is not, by itself, evidence of excess supply, and a change in registered unemployment need not be concentrated among graduates. Improvements or deterioration in individual matches can occur without the particular aggregate patterns considered here.

For an exploratory summary we also add the annual cross-provincial z-scores of the college-educated share and registered unemployment rate. This sum describes two relative provincial positions, not a supply–demand gap. In particular, a high value can be produced by a high education share, a high registered unemployment rate or both; it has no inherently adverse interpretation. This construction differs from formal mismatch measures that use information on the distribution of job seekers and vacancies across sectors or occupations (Şahin et al., 2014). We therefore report the sum alongside its separate components and do not interpret it as a measure of educational mismatch.

Institutional background

The National Big Data Comprehensive Pilot Zone programme began with the approval of Guizhou in February 2016. Seven further zones were approved in October 2016: Beijing–Tianjin–Hebei, the Pearl River Delta, Shanghai, Henan, Chongqing, Shenyang and Inner Mongolia (National Development and Reform Commission, 2020; State Information Center, 2016). The zones differ in geographic scope and combine infrastructure, administrative and industrial initiatives. Figure 3 shows the approvals and the annual treatment coding used in the analysis.

The Pearl River Delta covers only part of Guangdong, and Shenyang is one city in Liaoning. A province-level coding of either host province as treated assigns the same status to residents inside and outside the zone. The two geographic mappings therefore correspond to different samples as well as different treatment definitions; a difference between their coefficients does not by itself quantify the bias from misclassification.

Under the inclusive mapping, Guizhou is coded as treated from 2016 and the nine other host provinces from 2017, the first full year after the second approval. The strict mapping excludes Guangdong and Liaoning and retains eight treated units among 29 provinces. Figures 1 and 2 show the national education and digital-development context in which designation occurred; national trends alone do not identify the effect of the pilot programme (China Academy of Information and Communications Technology, 2023; China Internet Network Information Center, 2011, 2021).

Data and methods

Data sources and panel construction

The core panel covers all 31 mainland provincial-level units from 2010 to 2023, yielding 434 province-year observations. Hong Kong, Macao and Taiwan are excluded. Education and registered-unemployment data were transcribed from the National Bureau of Statistics' China Statistical Yearbook editions for 2011–2024 (National Bureau of Statistics of China, 2011–2024). No values were simulated, interpolated or imputed; outcome-specific samples are shorter where the relevant series is unavailable or comparability is restricted.

The 2010–2013 tables were parsed from official HTML-compatible or binary XLS files. For 2014–2023, values were extracted by structured transcription from image tables on the public yearbook pages. Deterministic checks assessed provincial coverage, numeric ranges, missing values and year-to-year discontinuities. Source images were spot-checked for years with layout or definition changes, including the 2015 education table, the 2020 census-based education table and the 2022–2023 unemployment tables.

The education outcome is the percentage of persons aged six and above with junior-college, university or postgraduate education. For province p in year t, it is calculated as:

CollegeShare□□ = 100 × (JuniorCollege□□ + University□□ + Postgraduate□□) / PopulationAged6Plus□□.

The 2020 education observation comes from the Seventh National Population Census, whereas observations in surrounding years come from annual population sample surveys; one sensitivity specification excludes 2020. The urban registered unemployment rate is available for 2010–2021 but is no longer published in the yearbook from 2022 onward. Registered unemployed persons are available through 2023, but an official statistical-scope adjustment from 2020 limits comparability with earlier years; the count regressions therefore end in 2019. Registered unemployment is an administrative measure and depends on registration, so it need not track survey unemployment (Giles et al., 2005; Feng et al., 2017). Table 1 summarises the series and their comparability limitations.

Table 1
Dataset elements and comparability issues

Dataset element	Coverage	Definition or unit	Key comparability issue
College-or-higher share	31 units, 2010–2023	Percentage of persons aged six and above	2020 census versus annual sample surveys
Urban registered unemployment rate	31 units, 2010–2021	Percentage points	Rate no longer published from 2022; 2020 scope change
Registered unemployed persons	31 units, 2010–2023	Persons; analysed in logarithms	Official scope change from 2020
Combined education–unemployment indicator (EMI)	31 units, 2010–2021	Sum of within-year z-scores (see below)	Inherits both source-series breaks; headline sample ends 2019
Inclusive treatment mapping	31 units	Guangdong and Liaoning coded treated from 2017	Partial-zone exposure treated as province-wide
Strict treatment mapping	29 units	Guangdong and Liaoning excluded	Smaller sample but less geographic misclassification

During 2010–2015, the strict-mapping treated provinces had an average college-or-higher share of 16.2%, compared with 10.1% in the never-treated provinces. Their mean registered unemployment rates were 3.28% and 3.45%, respectively. The education gap was driven in part by Beijing, Tianjin and Shanghai; differences in levels and trajectories are considered in the analyses below.

Exploratory combined relative-position indicator

We also examine an exploratory indicator, denoted EMI, that combines the province's relative position in the college-or-higher share with its relative position in registered unemployment in each year:

EMI□□ = z□(CollegeShare□□) + z□(UnempRate□□).

Here z□(·) standardises each variable across provinces within calendar year t. EMI is the sum of two within-year standardised scores, not a separately standardised score: a higher value means a higher combined relative position, and either component can dominate the sum. The indicator does not measure educational or worker–job mismatch, and neither an increase nor a decrease is a necessary condition for a change in individual matching. To show which component contributes to an estimated composite association, we estimate the same specifications separately for z□(CollegeShare) and z□(UnempRate). With the same observations and regressors, the two component coefficients sum to the EMI coefficient; this identity was checked computationally. The headline EMI sample is 2010–2019, before the 2020 changes in both series. A sensitivity specification extends the sample to 2021 while excluding 2020; the 2021 unemployment observation remains subject to the changed statistical scope.

Econometric specification

The province-level two-way fixed-effects (TWFE) specification compares changes in outcomes in designated and never-treated provinces. Under the inclusive mapping, Guizhou is coded as treated from 2016 and the nine other host provinces, including Guangdong and Liaoning, from 2017. The strict mapping excludes Guangdong and Liaoning in every year because the Pearl River Delta and Shenyang zones cover only parts of those provinces; it retains 29 provinces, of which eight are treated. The inclusive mapping retains all 31 provinces, of which ten are treated. A causal difference-in-differences interpretation requires the treated and comparison provinces to have followed parallel counterfactual trends in the absence of designation (Abadie, 2005). The estimating equation is

$Y_{pt} = \beta Policy_{pt} + \alpha_p + \lambda_t + \varepsilon_{pt}$.

where Y_{pt} is an outcome in province p and year t, $Policy_{pt}$ is the treatment indicator for the selected geographic mapping, and α_p and λ_t are province and year fixed effects. Standard errors are clustered by province (Bertrand et al., 2004), and P values are two-sided. Depending on the outcome, sensitivity specifications exclude 2020, add province-specific linear trends, restrict the analysis to pre-break years or omit Guizhou, the sole early-treated province. Year fixed effects absorb changes common to provinces in a given year; they do not by themselves separate a statistical-scope change from other contemporaneous influences that differ across provinces.

The conventional event study replaces the treatment indicator with indicators for time relative to treatment, using the year immediately before treatment as the reference. Relative years at or below − 4 and at or above + 5 are binned, and a joint Wald test assesses the pre-treatment coefficients. Because conventional TWFE event-study estimates can be affected by treatment-effect heterogeneity under staggered adoption (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021; Borusyak et al., 2024), we also estimate the education event study using the interaction-weighted approach of Sun and Abraham (2021). Cohort-by-relative-time indicators are estimated against never-treated provinces, with relative year − 1 omitted, and cohort-specific coefficients are aggregated using cohort shares. The interaction-weighted analysis uses the same endpoint bins and province-clustered standard errors. Guizhou is the only province in the 2016 cohort, limiting cohort-specific inference. The interaction-weighted approach addresses the heterogeneous-effect contamination of conventional TWFE estimates but still requires an appropriate counterfactual trend; failure to reject pre-treatment coefficients alone does not establish parallel trends (Roth et al., 2023).

Results

Descriptive trends

Under the strict mapping, the treated provinces' unweighted mean college-or-higher share exceeded that of never-treated provinces throughout the period (Fig. 4). The gap grew after 2016, but it had also

widened before designation. The treated-group registered unemployment rate fell relative to the control-group rate until the mid-2010s, then reversed and crossed above it after 2016. These descriptive differences motivate examination of pre-treatment trends as well as post-designation associations.

Education share estimates vary with province-specific trends

The 2010–2023 TWFE estimate for the college-or-higher share was 1.122 percentage points under the inclusive mapping (95% CI − 0.663 to 2.908; P = 0.209; n = 434) and 1.571 points under the strict mapping (95% CI − 0.353 to 3.494; P = 0.106; n = 406; Table 2). Excluding 2020 produced estimates of 1.256 points (95% CI − 0.641 to 3.154; P = 0.186) and 1.730 points (95% CI − 0.340 to 3.800; P = 0.098), respectively. With province-specific linear trends, the estimates were − 0.145 points (95% CI − 1.602 to 1.313; P = 0.841) and 0.268 points (95% CI − 1.208 to 1.743; P = 0.713). All confidence intervals included zero, and the estimates were sensitive to the trend specification; the intervals also allow effects in both directions.

Table 2
Two-way fixed-effects estimates for education and registered unemployment

Spec.	Outcome	Mapping, sample	Trends	Estimate	SE	95% CI	*P*	*n*
C1	College share	Incl., 2010–23	No	1.122	0.874	[− 0.663, 2.908]	0.209	434
C2	College share	Strict, 2010–23	No	1.571	0.939	[− 0.353, 3.494]	0.106	406
C3	College share	Incl., excl. 2020	No	1.256	0.929	[− 0.641, 3.154]	0.186	403
C4	College share	Strict, excl. 2020	No	1.730	1.011	[− 0.340, 3.800]	0.098	377
C5	College share	Incl., 2010–23	Yes	−0.145	0.714	[− 1.602, 1.313]	0.841	434
C6	College share	Strict, 2010–23	Yes	0.268	0.720	[− 1.208, 1.743]	0.713	406
U1	Unemp. rate	Incl., 2010–21	No	0.382	0.147	[0.083, 0.682]	0.014	372
U2	Unemp. rate	Strict, 2010–21	No	0.317	0.148	[0.013, 0.620]	0.042	348
U3	Unemp. rate	Incl., 2010–19	No	0.218	0.113	[− 0.014, 0.450]	0.064	310
U4	Unemp. rate	Strict, 2010–19	No	0.145	0.102	[− 0.063, 0.353]	0.164	290
N1	Log unemp. persons	Incl., 2010–19	No	0.019	0.050	[− 0.082, 0.120]	0.707	310
N2	Log unemp. persons	Strict, 2010–19	No	0.027	0.057	[− 0.089, 0.143]	0.637	290

Notes: ‘College share’ is the college-or-higher population share, with estimates in percentage points; ‘Unemp. rate’ is the urban registered unemployment rate, with estimates in percentage points; ‘Log unemp. persons’ is the natural logarithm of registered unemployed persons. ‘Trends’ indicates province-specific linear trends. All models include province and year fixed effects; standard errors (SE) are clustered by province. The aggregate outcomes do not measure individual overeducation.

Education event studies show pre-treatment differences

The education event studies show differences in pre-treatment trajectories between designated and comparison provinces (Fig. 5; Table 4). Joint tests of the conventional TWFE pre-treatment coefficients gave P = 0.048 under the inclusive mapping and P = 0.026 under the strict mapping. The corresponding interaction-weighted tests gave P = 0.001 and P = 0.006. The results raise concern about a causal

interpretation of the education estimates even when using an estimator designed for heterogeneous treatment effects. At relative year + 3, the conventional inclusive estimate was − 1.461 percentage points (95% CI − 2.835 to − 0.087); the strict estimate was − 1.309 points (95% CI − 2.898 to 0.281). Most other post-treatment coefficients were imprecise. Province-specific trends also substantially changed the pooled estimates, although they do not establish what the counterfactual trajectory would have been.

Registered unemployment associations differ across sample windows

In the 2010–2021 sample, designation was associated with a higher registered unemployment rate in the TWFE model: 0.382 percentage points under the inclusive mapping (95% CI 0.083 to 0.682; P = 0.014) and 0.317 points under the strict mapping (95% CI 0.013 to 0.620; P = 0.042; Table 2). These estimates are approximately 12% and 10%, respectively, of the strict-mapping treated provinces' pre-policy mean rate of 3.28%. The outcome reflects administrative registration and does not by itself establish a change in overall unemployment or a causal effect of designation.

Estimates were smaller when the sample ended in 2019, before the documented scope adjustment: 0.218 percentage points under the inclusive mapping (95% CI − 0.014 to 0.450; P = 0.064) and 0.145 points under the strict mapping (95% CI − 0.063 to 0.353; P = 0.164). For the logarithm of registered unemployed persons in 2010–2019, the estimates were 0.019 (95% CI − 0.082 to 0.120; P = 0.707) and 0.027 (95% CI − 0.089 to 0.143; P = 0.637), respectively. Comparing these windows does not identify how much of the difference is attributable to the 2020 scope change: the longer sample also includes later years and other contemporaneous developments, while year fixed effects absorb common annual shifts. The intervals for the pre-break estimates include positive and negative associations.

The exploratory composite is dominated by the unemployment component

In 2010–2019, the estimated association with EMI was 0.342 combined-score units under the inclusive mapping (95% CI − 0.041 to 0.724; P = 0.078; n = 310) and 0.273 under the strict mapping (95% CI − 0.119 to 0.665; P = 0.165; n = 290; Table 3). Extending the sample to 2021 while excluding 2020 gave estimates of 0.521 (95% CI 0.087 to 0.955; P = 0.020) and 0.455 (95% CI − 0.012 to 0.921; P = 0.056), respectively. This extended window includes post-adjustment unemployment data. With province-specific trends, the 2010–2019 estimates were − 0.134 (95% CI − 0.543 to 0.276; P = 0.511) and − 0.156 (95% CI − 0.586 to 0.275; P = 0.466). Omitting Guizhou gave estimates of 0.406 (95% CI 0.020 to 0.792; P = 0.040) and 0.337 (95% CI − 0.069 to 0.742; P = 0.101). Six of the eight point estimates were positive; most confidence intervals included zero. A positive estimate indicates a higher combined relative position, not worse educational matching.

The positive pre-break composite estimates were concentrated in the standardised registered-unemployment component. Under the inclusive mapping, the 0.342 composite coefficient was the sum

of − 0.009 for the standardised college share (P = 0.910) and 0.351 for standardised registered unemployment (P = 0.070). Under the strict mapping, the corresponding 0.273 coefficient comprised 0.040 (P = 0.576) and 0.233 (P = 0.160). Across all eight matched specifications, the standardised college-share coefficients were small and none was statistically different from zero (P ≥ 0.188), while the unemployment component accounted for nearly all positive estimates in models without province-specific trends. The component coefficients add to the composite coefficient to numerical precision; their imprecision does not establish that the education component is exactly zero.

The EMI event study yielded joint pre-treatment P values of 0.087 under the inclusive mapping and 0.188 under the strict mapping (Fig. 6, left). Post-treatment coefficients were individually imprecise. The pre-treatment tests do not reject at the 5% level, but they do not establish parallel counterfactual trends. Figure 6, right, displays the sensitivity of the composite estimates to the specification; none of these aggregate comparisons directly tests changes in worker–job matching.

Table 3
Two-way fixed-effects estimates for the exploratory combined relative-position indicator (EMI)

Spec.	Mapping	Sample	Trends	Estimate	SE	95% CI	*P*	*n*
M1	Inclusive	2010–2019	No	0.342	0.187	[− 0.041, 0.724]	0.078	310
M2	Strict	2010–2019	No	0.273	0.192	[− 0.119, 0.665]	0.165	290
M3	Inclusive	2010–2021, excl. 2020	No	0.521	0.213	[0.087, 0.955]	0.020	341
M4	Strict	2010–2021, excl. 2020	No	0.455	0.228	[− 0.012, 0.921]	0.056	319
M5	Inclusive	2010–2019	Yes	−0.134	0.201	[− 0.543, 0.276]	0.511	310
M6	Strict	2010–2019	Yes	−0.156	0.210	[− 0.586, 0.275]	0.466	290
M7	Inclusive	2010–2019, excl. Guizhou	No	0.406	0.189	[0.020, 0.792]	0.040	300
M8	Strict	2010–2019, excl. Guizhou	No	0.337	0.198	[− 0.069, 0.742]	0.101	280

Notes: EMI is the within-year z-score of the college-or-higher share plus the within-year z-score of the registered unemployment rate. Each component is measured in its own within-year standard-deviation units; their sum is not separately standardised. All models include province and year fixed effects; standard errors (SE) are clustered by province; ‘Trends’ indicates province-specific linear trends. EMI is an exploratory combined relative-position indicator, not a measure of educational mismatch or individual overeducation.

Table 4
Event-study coefficients for the college-or-higher share

Relative year	Inclusive estimate (95% CI)	Strict estimate (95% CI)
≤−4	−2.182 [− 4.206, − 0.157]	−2.432 [− 4.826, − 0.038]
−3	−1.227 [− 3.470, 1.015]	−1.174 [− 3.873, 1.525]
−2	−1.404 [− 3.174, 0.366]	−1.220 [− 3.262, 0.822]
−1	0 (reference)	0 (reference)
0	−0.383 [− 1.535, 0.770]	−0.047 [− 1.288, 1.194]
+ 1	−0.380 [− 1.743, 0.983]	0.270 [− 0.798, 1.339]
+ 2	0.162 [− 1.130, 1.454]	0.655 [− 0.587, 1.897]
+ 3	−1.461 [− 2.835, − 0.087]	−1.309 [− 2.898, 0.281]
+ 4	−0.448 [− 1.748, 0.851]	−0.189 [− 1.612, 1.235]
≥+5	−0.400 [− 1.749, 0.949]	−0.134 [− 1.559, 1.290]

Notes: Outcome is the college-or-higher share in percentage points; the omitted category is the year immediately before treatment. Pre-treatment joint tests: inclusive P = 0.048; strict P = 0.026.

Discussion

Interpreting the provincial associations

The observed provincial associations do not answer whether the big-data zones improved workers' educational matches. The more defensible finding is narrower: registered unemployment was higher in pilot provinces relative to comparison provinces in the 2010–2021 fixed-effects specifications, while estimated education associations depend on the treatment of pre-existing trends. These patterns combine changes in the underlying labour market with changes in who resides in a province and who registers as unemployed.

Registration is particularly important for interpretation. The registered rate is an administrative measure, not an estimate of unemployment among all workers or among graduates; its coverage and movement differ from those of survey measures (Giles et al., 2005; Feng et al., 2017). Digital public services could affect access to registration, while job loss or job creation could also change the number of people seeking registration. The available data do not distinguish these explanations. The smaller 2010–2019 estimates and the larger 2010–2021 estimates also compare different periods, so the difference cannot be attributed uniquely to the 2020 statistical change.

The education event studies add a separate warning about the counterfactual. Pre-treatment coefficients are jointly different from zero in both the conventional and interaction-weighted specifications. Although a test of pre-trends cannot establish the counterfactual path, the rejection is evidence against the restrictions needed for a simple causal reading of the education estimates. Province-specific trends substantially change the coefficients, but adding trends does not establish that the remaining comparison is unconfounded.

Comparison with other digital-policy studies

Ling et al. (2023) report that the smart-city pilot programme reduced employment pressure in Chinese cities. Our positive full-period estimates concern registered unemployment in provinces exposed to a different programme, and the estimates are less precise before 2020. These differences in policy, geography, outcome and period offer plausible reasons for non-comparable patterns; the present data do not isolate which difference matters. Li and Li (2025) study employment at the firm level under the same big-data-zone programme, while Xu et al. (2024) report a rise in firms' labour-income share. Neither outcome is a provincial registered unemployment rate. At the individual level, Cao and Cao (2024) associate digital skills with lower education–job mismatch; our province-level series cannot determine whether that relationship held for workers in the pilot zones.

The inclusive and strict geographic specifications also answer slightly different questions. Including Guangdong and Liaoning preserves observations but treats their partial zones as if they covered the whole province; excluding both avoids that coding choice but changes the sample and target population. Similarly, ending the unemployment series in 2019 avoids the documented 2020 comparability issue at the cost of discarding later observations. Neither comparison identifies the size of a single measurement bias, but each shows how the reported association changes with an explicit analytical choice.

Limitations and implications

The strongest limitation is measurement: neither schooling-to-job requirements nor unemployment by educational attainment is observed. The exploratory sum of standardised education and unemployment measures does not fill that gap; its positive non-trend coefficients are largely attributable to the registered-unemployment component. Causal interpretation is also limited by divergent education pre-trends, overlapping policies, a small number of treated provinces and a single early-treated province. Survey and census differences in the education series and the change in unemployment-series scope further complicate comparisons over time. Matched worker–job data and outcomes by education would be needed to study overeducation directly, while finer geographic exposure data would distinguish the partial zones from the rest of their host provinces.

Conclusion

The National Big Data Comprehensive Pilot Zones were followed by higher relative registered-unemployment estimates in the full 2010–2021 provincial sample, but smaller and less precise

estimates in 2010–2019. Education estimates are sensitive to pre-existing trajectories. These findings describe changes in aggregate indicators around policy designation; they do not establish whether designation caused the changes or whether the education–job match improved.

Evaluations of digital policy should therefore distinguish provincial administrative indicators from the employment experiences of individual workers and should report the geography and comparability of each data series. A direct assessment of overeducation requires observations linking workers' qualifications to their jobs, together with a design able to account for pre-existing differences between pilot and comparison areas.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study used publicly available aggregate provincial statistics and did not involve identifiable individual participants or animals.

Informed consent

This study did not use identifiable individual participant data.

References

1. Abadie A (2005) Semiparametric difference-in-differences estimators. Rev Econ Stud 72(1):1–19. https://doi.org/10.1111/0034-6527.00321
2. Acemoglu D, Autor D (2011) Skills, tasks and technologies: implications for employment and earnings. In: Ashenfelter O, Card D (eds) Handbook of labor economics, vol 4B. Elsevier, Amsterdam, pp 1043–1171
3. Acemoglu D, Restrepo P (2020) Robots and jobs: evidence from US labor markets. J Polit Econ 128(6):2188–2244. https://doi.org/10.1086/705716
4. Allen J, van der Velden R (2001) Educational mismatches versus skill mismatches: effects on wages, job satisfaction, and on-the-job search. Oxf Econ Pap 53(3):434–452. https://doi.org/10.1093/oep/53.3.434
5. Autor DH (2001) Wiring the labor market. J Econ Perspect 15(1):25–40. https://doi.org/10.1257/jep.15.1.25
6. Autor DH (2015) Why are there still so many jobs? The history and future of workplace automation. J Econ Perspect 29(3):3–30. https://doi.org/10.1257/jep.29.3.3

7. Autor DH, Levy F, Murnane RJ (2003) The skill content of recent technological change: an empirical exploration. Q J Econ 118(4):1279–1333. https://doi.org/10.1162/003355303322552801
8. Bai L (2006) Graduate unemployment: dilemmas and challenges in China's move to mass higher education. China Q 185:128–144. https://doi.org/10.1017/S0305741006000087
9. Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? Q J Econ 119(1):249–275. https://doi.org/10.1162/003355304772839588
10. Borusyak K, Jaravel X, Spiess J (2024) Revisiting event-study designs: robust and efficient estimation. Rev Econ Stud 91(6):3253–3285. https://doi.org/10.1093/restud/rdae007
11. Callaway B, Sant'Anna PHC (2021) Difference-in-differences with multiple time periods. J Econom 225(2):200–230. https://doi.org/10.1016/j.jeconom.2020.12.001
12. Cao H, Cao L (2024) The effect of digital skills on education–job mismatch: evidence from the China Family Panel Studies. Labor Econ Res 12(1):102–122. http://iple.cass.cn/dqcbw/ldjjyj/qkml/202405/t20240523_5754204.shtml (in Chinese)
13. Chevalier A (2003) Measuring over-education. Economica 70(279):509–531. https://doi.org/10.1111/1468-0335.t01-1-00296
14. China Academy of Information and Communications Technology (2023) Research report on the development of China's digital economy (2023). CAICT, Beijing. http://www.caict.ac.cn/kxyj/qwfb/bps/
15. China Internet Network Information Center, Beijing (2011) The 27th statistical report on China's internet development. CNNIC. https://www.cnnic.com.cn/IDR/
16. China Internet Network Information Center, Beijing (2021) The 47th statistical report on China's internet development. CNNIC. https://www.cnnic.com.cn/IDR/
17. de Chaisemartin C, D'Haultfœuille X (2020) Two-way fixed effects estimators with heterogeneous treatment effects. Am Econ Rev 110(9):2964–2996. https://doi.org/10.1257/aer.20181169
18. Deming DJ (2017) The growing importance of social skills in the labor market. Q J Econ 132(4):1593–1640. https://doi.org/10.1093/qje/qjx022
19. Duncan GJ, Hoffman SD (1981) The incidence and wage effects of overeducation. Econ Educ Rev 1(1):75–86. https://doi.org/10.1016/0272-7757(81)90028-5
20. Feng S, Hu Y, Moffitt R (2017) Long run trends in unemployment and labor force participation in urban China. J Comp Econ 45(2):304–324. https://doi.org/10.1016/j.jce.2017.02.004
21. Giles J, Park A, Zhang J (2005) What is China's true unemployment rate? China Econ Rev 16(2):149–170. https://doi.org/10.1016/j.chieco.2004.11.002
22. Goodman-Bacon A (2021) Difference-in-differences with variation in treatment timing. J Econom 225(2):254–277. https://doi.org/10.1016/j.jeconom.2021.03.014
23. Goos M, Manning A, Salomons A (2014) Explaining job polarization: routine-biased technological change and offshoring. Am Econ Rev 104(8):2509–2526. https://doi.org/10.1257/aer.104.8.2509

24. Kuhn P, Mansour H (2014) Is internet job search still ineffective? Econ J 124(581):1213–1233. https://doi.org/10.1111/ecoj.12119
25. Leuven E, Oosterbeek H (2011) Overeducation and mismatch in the labor market. In: Hanushek EA, Machin S, Woessmann L (eds) Handbook of the economics of education, vol 4. Elsevier, Amsterdam, pp 283–326
26. Li M, Li H (2025) Beijing Soc Sci 3:76–90. https://doi.org/10.13262/j.bjsshkxy.bjshkx.250307(in Chinese) The employment promotion effect of big data development: a quasi-natural experiment based on the National Big Data Comprehensive Pilot Zone
27. Li S, Whalley J, Xing C (2014) China's higher education expansion and unemployment of college graduates. China Econ Rev 30:567–582. https://doi.org/10.1016/j.chieco.2013.08.002
28. Ling X, Luo Z, Feng Y, Liu X, Gao Y (2023) How does digital transformation relieve the employment pressure in China? Empirical evidence from the national smart city pilot policy. Humanit Soc Sci Commun 10:617. https://doi.org/10.1057/s41599-023-02131-w
29. McGuinness S (2006) Overeducation in the labour market. J Econ Surv 20(3):387–418. https://doi.org/10.1111/j.0950-0804.2006.00284.x
30. Ministry of Education (2023) The Ministry of Education and the Ministry of Human Resources and Social Security jointly deploy work on employment and entrepreneurship for the 2024 graduating cohort. Ministry of Education, Beijing. http://www.moe.gov.cn/jyb_zzjg/huodong/202312/t20231205_1093287.html
31. National Bureau of Statistics of China (2011–2024) China statistical yearbook. China Statistics, Beijing. https://www.stats.gov.cn/sj/ndsj/
32. National Development and Reform Commission (2020) National Big Data (Guizhou) Comprehensive Pilot Zone carries out innovative experiments. NDRC, Beijing. https://www.ndrc.gov.cn/xwdt/ztzl/szhzxhbxd/zxal/202007/t20200703_1233042.html
33. Roth J, Sant'Anna PHC, Bilinski A, Poe J (2023) What's trending in difference-in-differences? A synthesis of the recent econometrics literature. J Econom 235(2):2218–2244. https://doi.org/10.1016/j.jeconom.2023.03.008
34. Şahin A, Song J, Topa G, Violante GL (2014) Mismatch unemployment. Am Econ Rev 104(11):3529–3564. https://doi.org/10.1257/aer.104.11.3529
35. Spitz-Oener A (2006) Technical change, job tasks, and rising educational demands: looking outside the wage structure. J Labor Econ 24(2):235–270. https://doi.org/10.1086/499972
36. The second batch of seven National Big Data Comprehensive Pilot Zones was approved. SIC, State Information Center, Beijing (2016) https://scdrc.sic.gov.cn/SmarterCity_new/szjj/638/0408/20250408111249343149280_mobile.html
37. Sun J, Zu N, Zhang C (2025) Big data and firm greenwashing: evidence from China's National Big Data Comprehensive Pilot Zone. Humanit Soc Sci Commun 12:768. https://doi.org/10.1057/s41599-025-05107-0

38. Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J Econom 225(2):175–199. https://doi.org/10.1016/j.jeconom.2020.09.006
39. World Bank (2024) School enrollment, tertiary (% gross)—China, indicator SE.TER.ENRR. World Development Indicators. https://data.worldbank.org/indicator/SE.TER.ENRR?locations=CN
40. World Bank Group (2016) World development report 2016: digital dividends. World Bank. https://doi.org/10.1596/978-1-4648-0671-1. Washington, DC
41. Xiong B, Sui Q (2025) The effect of digital economy on rural workforce occupation transformation ability: evidence from China. Humanit Soc Sci Commun 12:13. https://doi.org/10.1057/s41599-024-04326-1
42. Xu Y, Wei Y, Zeng X, Yu H, Chen H (2024) Econ Anal Policy 84:1415–1437. https://doi.org/10.1016/j.eap.2024.10.031. Big data development and labor income share: evidence from China's National Big Data Comprehensive Pilot Zones

Figures

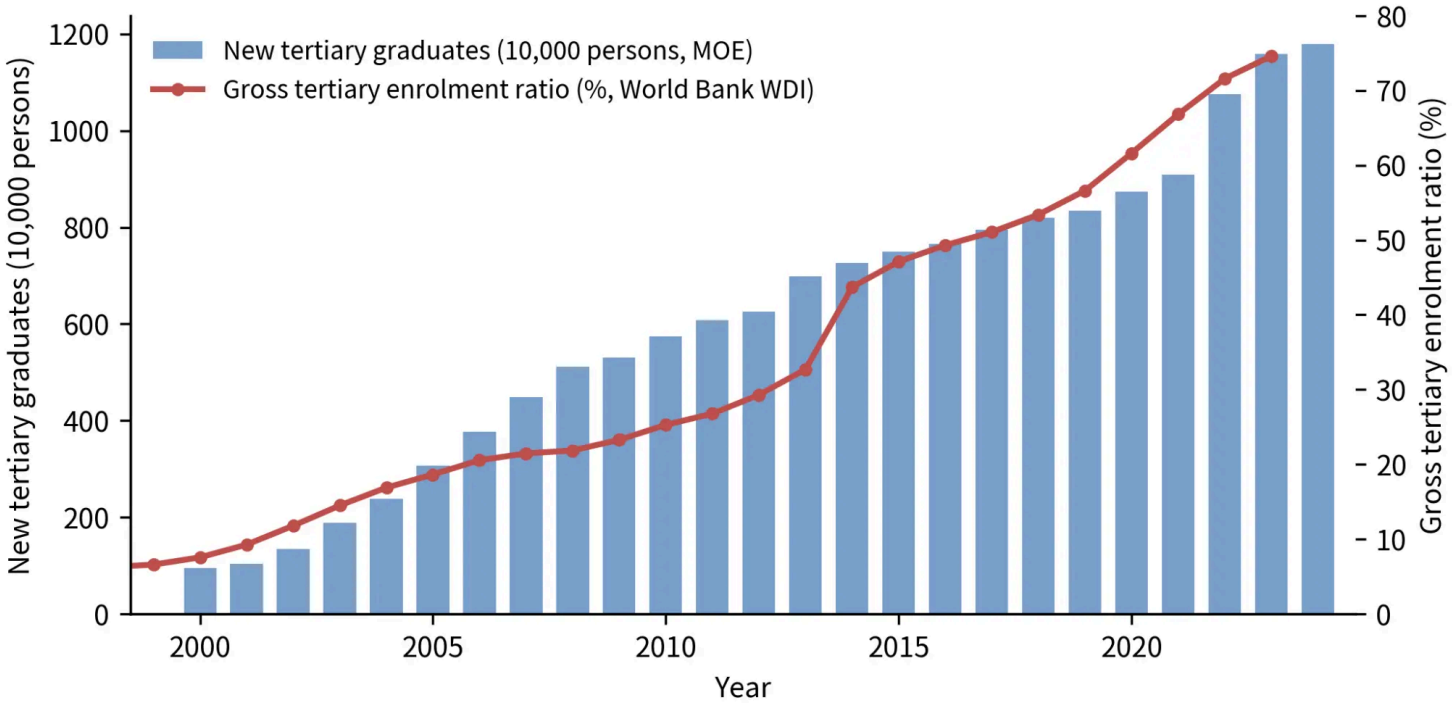


Figure 1

Higher-education expansion in China, 1999–2024. Bars show the reported size of annual tertiary-graduate cohorts in units of 10,000 persons; the 2024 cohort figure of 11.79 million was an advance estimate, not a realised count (Ministry of Education, 2023). The line shows the World Bank/UNESCO Institute for Statistics gross tertiary enrolment ratio (World Bank, 2024); this international measure is not identical to the domestic Ministry of Education ratio.

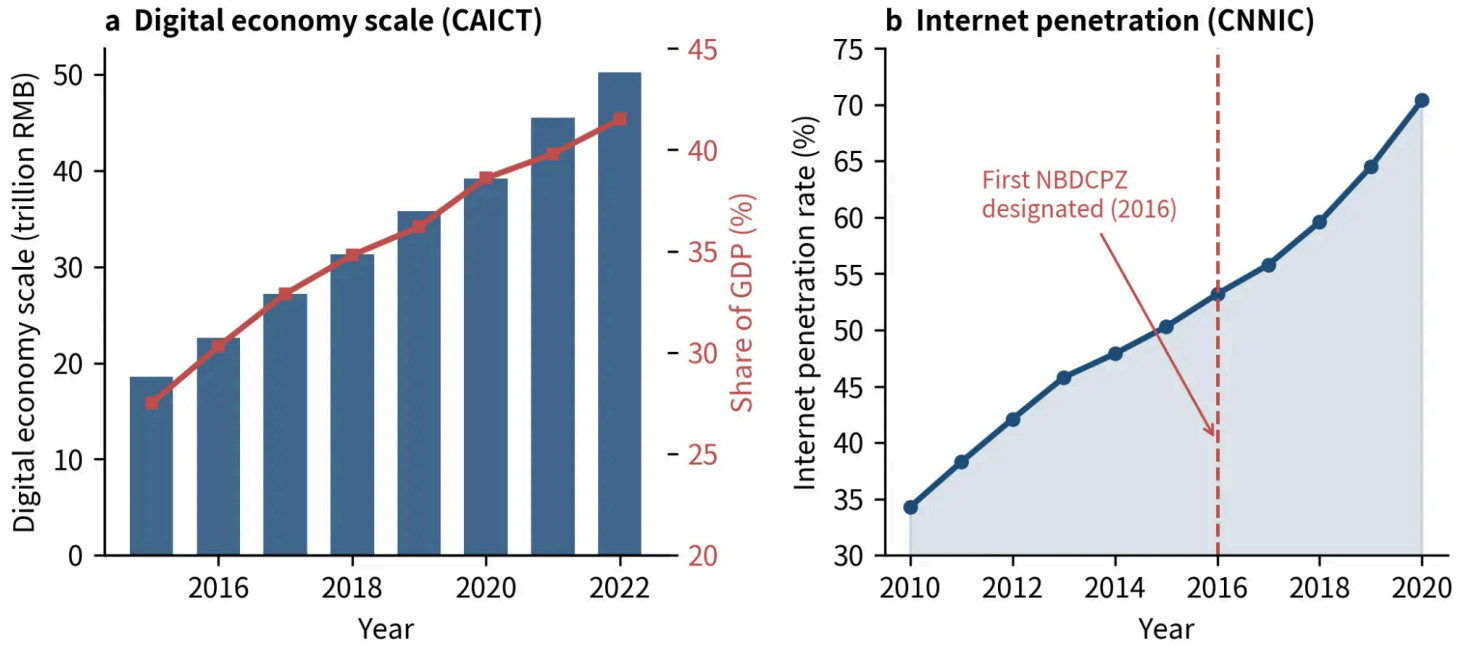
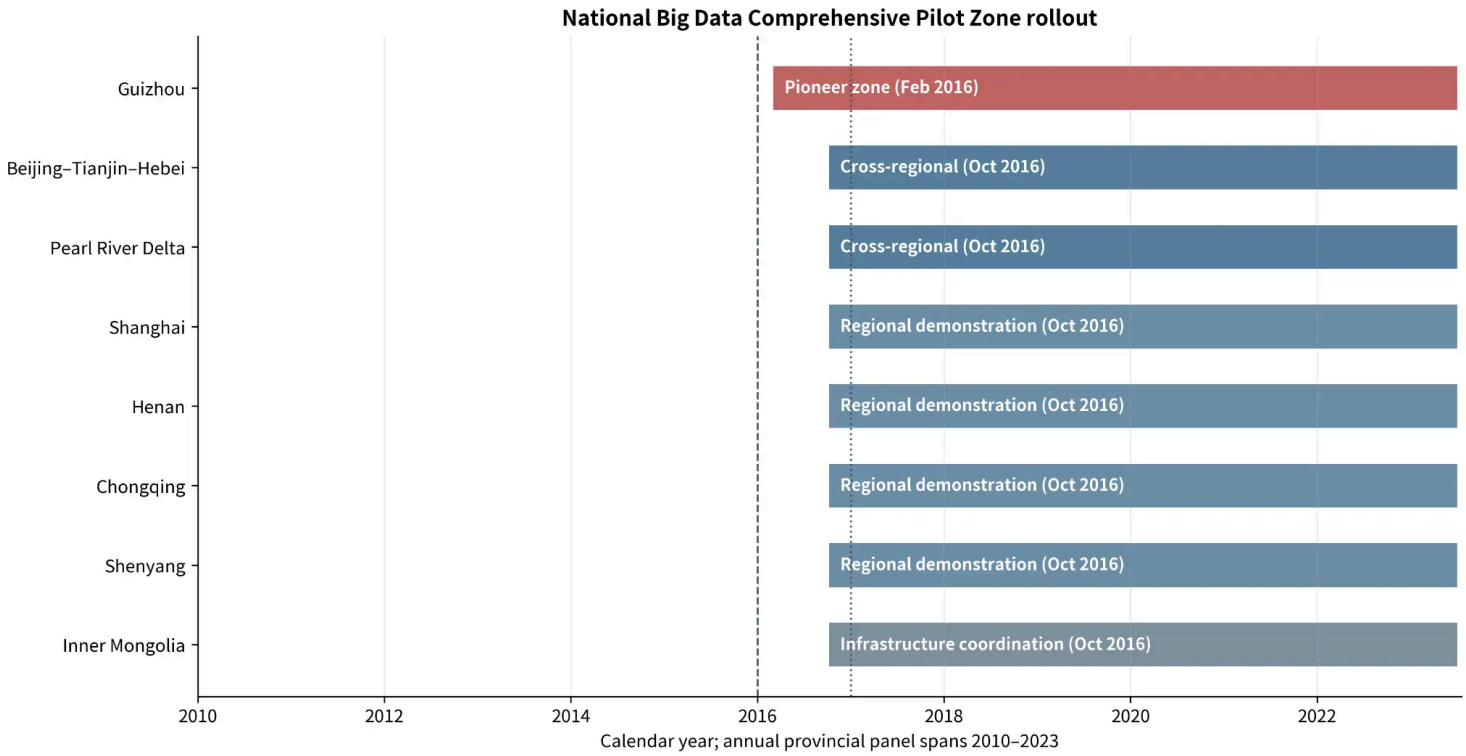


Figure 2

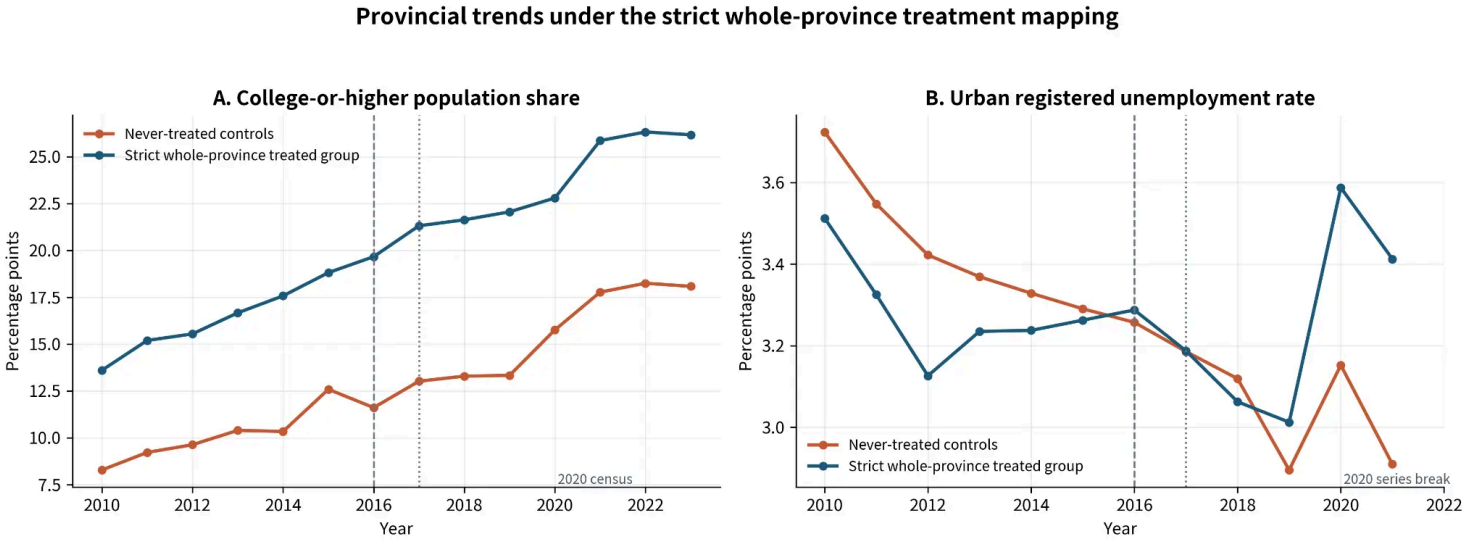
National digital-economy and internet-use trends. a, Estimated digital-economy scale and share of GDP, 2015–2022, from the China Academy of Information and Communications Technology. b, Internet penetration, 2010–2020; the 2010 and 2020 values are reported in the 27th and 47th China Internet Network Information Center reports. The vertical line marks the first pilot-zone designation in 2016.



Notes: Official geographic units are shown. The Pearl River Delta is not all Guangdong, and Shenyang is not all Liaoning. Annual treatment coding starts in 2016 for Guizhou and 2017 for the October 2016 batch.

Figure 3

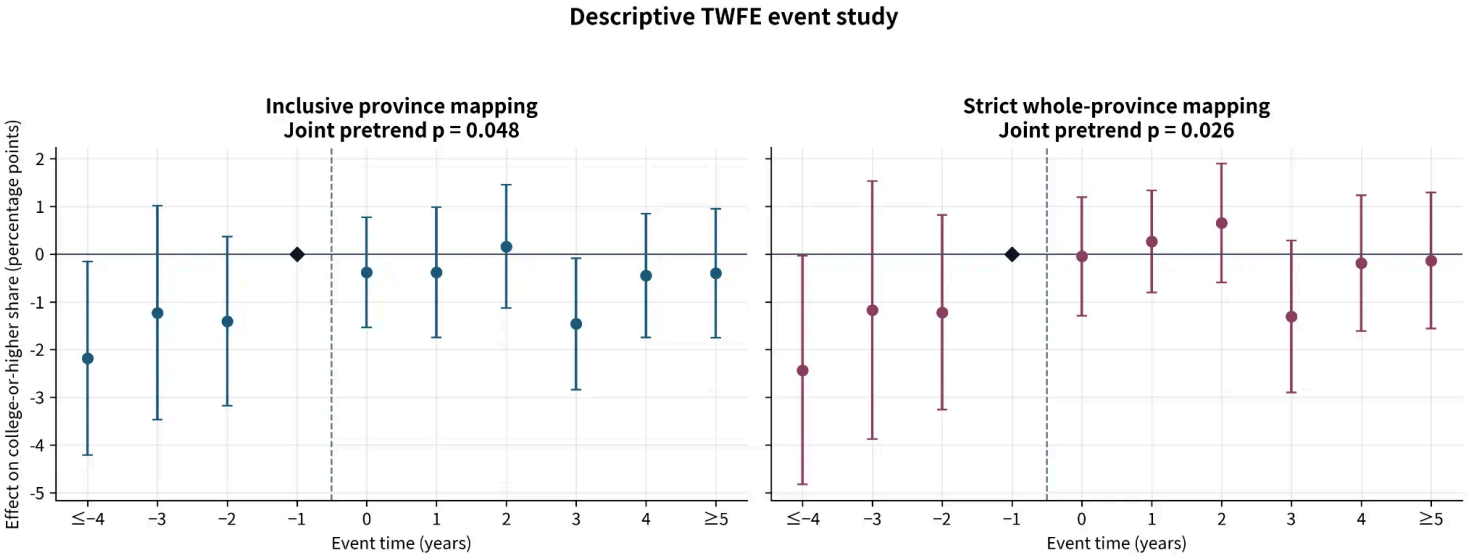
Rollout of the National Big Data Comprehensive Pilot Zones. The eight officially designated geographic units are shown. Guizhou was approved in February 2016 and is coded as treated from 2016; the second batch was approved in October 2016 and is coded as treated from 2017. The Pearl River Delta and Shenyang cover only parts of their host provinces.



Notes: Unweighted province means. Guangdong and Liaoning are excluded because official coverage is the Pearl River Delta and Shenyang only. Dashed/dotted lines mark 2016/2017 treatment timing.

Figure 4

Provincial outcome trends under the strict treatment mapping. Lines show unweighted annual means for treated and never-treated provincial-level units; Guangdong and Liaoning are excluded. The vertical markers denote the 2016 and 2017 treatment starts.



Notes: Province and year fixed effects; province-clustered 95% confidence intervals. Event time −1 is the omitted reference. Endpoints pool ≤−4 and ≥5. The design is descriptive and not cohort-robust.

Figure 5

Conventional two-way fixed-effects event-study estimates for the college-or-higher population share. Points show relative-year coefficients in percentage points; whiskers are 95% confidence intervals based on standard errors clustered by province. The year immediately before treatment is the reference; relative years ≤−4 and ≥+5 are pooled. The joint tests of pre-treatment coefficients yield P = 0.048 for the inclusive mapping and P = 0.026 for the strict mapping.

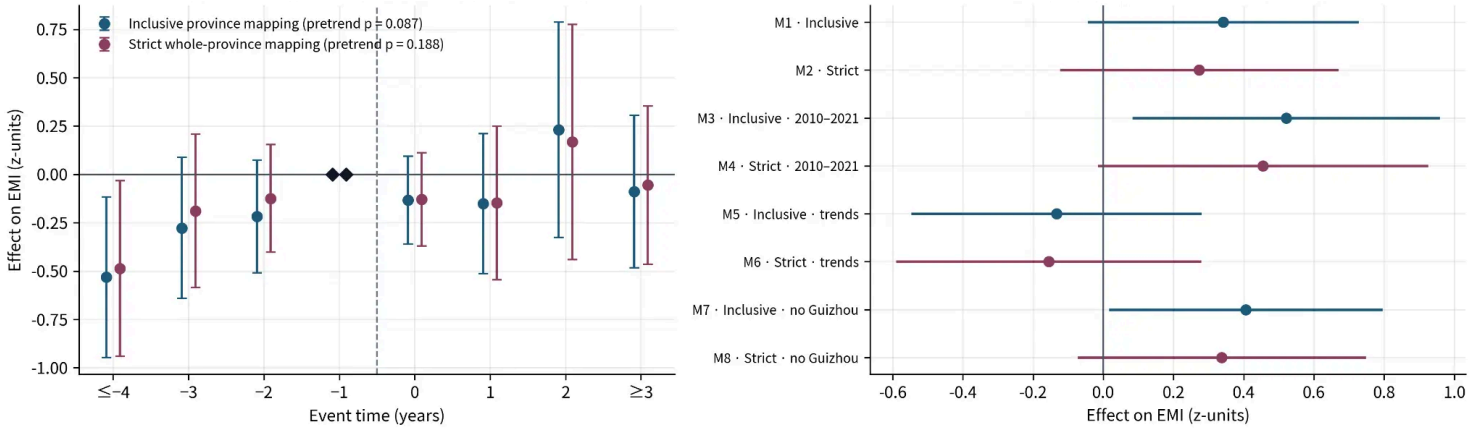


Figure 6

Exploratory combined education–unemployment indicator: event study and sensitivity. Left, event-study coefficients for 2010–2019 under both treatment mappings; whiskers show province-clustered 95% confidence intervals. Joint pre-treatment tests yield P = 0.087 for the inclusive mapping and P = 0.188 for the strict mapping. Right, estimates and 95% confidence intervals for the specifications in Table 3. The indicator, labelled EMI, is the sum of the within-year provincial z-scores for education share and registered unemployment; it is not a measure of worker–job mismatch.