We thank the reviewer for their careful and constructive comments. We have considered each point and outline below how we plan to address the comments in the revised manuscript.

Referees comments:

1. The causal claims are not convincing. This is a cross-sectional survey collected four to five years after the outbreak. IPTW cannot establish the timing of business actions and recovery. Most reported effects should be called associations.

Response:

Because the data are cross-sectional and do not establish the temporal ordering of support receipt, business actions, and recovery, the manuscript will be reframed as an observational analysis of covariate-adjusted associations. Terms such as “effect,” “contribution,” “effectiveness,” and “determinant” will be replaced throughout the Abstract, Methods, Results, Discussion, figure captions, and policy implications with “association,” “adjusted difference,” or equivalent non-causal wording. The estimates will be described as ATT-form doubly robust adjusted contrasts rather than causal ATT estimates. We will also state explicitly that the doubly robust property does not address unmeasured confounding, reverse causality, survivor selection, or ambiguous temporal ordering.

2. The covariate strategy is inconsistent. Revenue loss and employment change may occur after loans, subsidies, or adaptation decisions. Controlling for them could remove part of the treatment effect or introduce collider bias. The authors need treatment-specific causal diagrams and clearly justified covariate sets.

Response：

We understand the concern that revenue loss and employment change may be post-treatment variables for some loan, subsidy, and adaptation decisions, and that controlling for them may result in overadjustment or collider bias. In the revision, we will clarify the assumed temporal ordering and provide treatment-specific causal

diagrams and a transparent justification of the covariate set used for each treatment. Model B, which excludes revenue loss and employment change from both the propensity-score and outcome models, will be used as the main specification. The original Model A specification will be reported as a sensitivity analysis in the Supplement.

3.The propensity-score diagnostics are inadequate. Where are the propensity-score distributions, common-support checks, extreme weights, weight truncation, and effective sample sizes? Showing only selected pooled balance results is not sufficient for 26 treatments and numerous subgroup analyses.

Response:

We understand the reviewer's request for more comprehensive propensity-score and weighting diagnostics. To address this, the revision will expand the diagnostics currently reported. For all 26 pooled specifications, the Supplement will report propensity-score distributions by treatment status, empirical common support, the number and proportion of observations outside support, weight summaries including upper percentiles and maxima, effective sample sizes, and covariate-specific standardized mean differences before and after weighting. For each regional and sectoral estimate retained in the analysis, a diagnostic table will report the raw treated and control sample sizes, effective sample size, overlap, weight concentration, and maximum post-weighting absolute SMD. Alternative weight-truncation rules will be examined as sensitivity analyses. Estimates with poor overlap, extreme weights, or unresolved imbalance will not be substantively interpreted and may be removed from the main figures.

4. Equation 4 appears incorrect. The denominator seems to use the difference between the two variances rather than their pooled sum. The authors must verify the equation and confirm what was implemented in the code.

Response:

Equation (4) will be corrected so that the denominator uses the pooled sum of the two group variances rather than their difference. We will audit the analysis code and state explicitly which formula was implemented. If the code used the correct pooled-variance formula, the issue will be identified as a typographical error in the manuscript. If the implementation was also incorrect, all balance diagnostics will be recomputed, Figure 5 and the related tables will be updated, and any resulting changes in the analysis will be reported.

5.The manuscript performs a very large number of statistical tests. There are 26 treatments, followed by pooled, regional, and sectoral analyses. No correction for multiple comparisons is reported. How many “significant” findings would remain after controlling the false discovery rate?

Response:

We understand the concern regarding multiple comparisons. To address this, the revision will apply the Benjamini–Hochberg procedure at a 5% false discovery rate to the Model B results. The 26 pooled estimates will be treated as the primary family of hypotheses, while the regional and sectoral estimates will be treated as separate secondary families. We will report both raw and FDR-adjusted p-values and the number of findings that remain significant after correction. A more conservative correction across all pooled, regional, and sectoral estimates will also be examined as a sensitivity analysis.

6.The uncertainty estimation is unclear. How were the 95% confidence intervals calculated? Were propensity scores re-estimated during bootstrapping? Were standard errors clustered by country, city, or sector?

Response:

The uncertainty-estimation procedure will be described explicitly. For the revised primary estimates, 95% confidence intervals will be obtained using at least 1,000 nonparametric bootstrap replications. The propensity-score model, common-support assessment, weight-truncation procedure, and outcome model will be re-

estimated in every bootstrap sample. Resampling will be stratified by case region to preserve the case composition of the pooled dataset and will be conducted at the firm level unless the survey documentation identifies a higher-level recruitment or sampling unit. Any clustered sensitivity analysis will follow the actual sampling design; country, city, or sector will not be treated as a cluster mechanically, and the selected resampling unit will be justified in the Methods.

7.The threshold of 30 treated and control observations is arbitrary. Raw sample size does not guarantee reliable IPTW estimation. A subgroup with 30 observations and poor overlap may provide a meaningless estimate.

Response:

We understand the reviewer's concern regarding reliance on the threshold of at least 30 treated and 30 control observations. To address this, the threshold may be retained as an initial feasibility screen, but the reliability of subgroup estimates will be assessed more comprehensively using empirical overlap, effective sample size, weight concentration, and post-weighting covariate balance. Raw treated and control sample sizes will continue to be reported for transparency. Estimates with inadequate common support, very low effective sample size, highly concentrated weights, or unresolved balance problems will be omitted from the main results or identified as requiring cautious interpretation. The diagnostic information supporting each retained subgroup estimate will be provided in the Supplement.

8.The survey methodology is seriously underreported. Sampling frames, recruitment procedures, response rates, nonresponse adjustments, country-specific survey dates, and harmonization procedures are missing. Referring readers to a website that is still “under preparation” is unacceptable.

Response:

The survey methodology will be expanded substantially in the manuscript and Supplement. A case-specific table will document the sampling frame, target population and eligibility criteria, recruitment procedure, country- or case-specific

survey dates, survey mode, number invited or contacted, number starting and completing the survey where available, response- or completion-rate definition, nonresponse adjustment, incentives, survey language, translation procedures, questionnaire deviations, harmonization rules, ethics approval, and final analytic exclusions. Where the recruitment denominator is unavailable, a conventional response rate will not be inferred; the absence of the denominator and its implications for representativeness will be stated explicitly. DesignSafe will be used as a supporting archive, not as a substitute for essential methodological information in the paper.

9.How were permanently closed businesses handled? Surveying only surviving firms creates major survivor bias. A study of business recovery cannot ignore firms that failed completely.

Response:

The revision will clarify whether permanently closed businesses were eligible for or identifiable in each regional sampling frame and will report any available contact outcomes and confirmed closure counts. Where complete closure information is unavailable, survivor selection cannot be removed statistically. The target population and all conclusions will therefore be defined explicitly as businesses that were operating, reachable, and responding at the time of the 2024–2025 survey, rather than the full population of firms operating before the pandemic. Unobserved closed firms will not be assigned a recovery value of zero without supporting data. Survivor bias will be treated as a central limitation, and the findings will not be generalized to firms that failed completely. Available closure information will be reported descriptively by case.

10. The outcome is a retrospective, self-reported sales percentage. The authors provide no assessment of recall error, outliers, skewness, or cross-country measurement comparability. A simple linear outcome model requires much stronger diagnostics and sensitivity analyses.

Response:

The revision will add outcome diagnostics overall and by case region, including missingness, means and medians, dispersion, selected percentiles, skewness, heaping at common response values, extreme observations, and regression residual and influence diagnostics. Sensitivity analyses will re-estimate the main associations using a winsorized continuous outcome, a log sales-recovery ratio, a binary indicator of recovery to at least 100% of pre-pandemic sales, and a within-case standardized or ranked outcome to assess cross-country scale differences. We will also document the common question wording, reference period, translation, and harmonization procedures across cases. Because administrative sales records are not available for validation, recall error and cross-country reporting differences will be acknowledged explicitly and the conclusions will be limited accordingly.