Journal of International Business Studies
https://doi.org/10.1057/s41267-024-00725-3

EDITORIAL

Using the difference-in-differences design with panel data in international business research: progress, potential issues, and practical suggestions

Jiatao Li[1] · Han Jiang[2] · Jia Shen[3] · Haoyuan Ding[4] · Rongjian Yu[5]

Difference-in-differences (DID) is a popular design of causal inference in social science research. The rationale is to identify a particular event (the treatment) that influences some subjects (the treated group) but not others (the control group) and compare the differences in an outcome of interest between the treated and control groups before and after the treatment. This design constitutes a quasi-experiment that helps rule out confounding effects. As such, when applied with panel data (or, time-series cross-sectional data), DID design offers robust causal inference in terms of Granger causality between the time series of the treatment and that of the outcome of interest.

This rationale of DID design aligns well with the nature of international business (hereafter, IB) research. Firms' IB practices are jointly shaped by location-specific factors in different countries and firm- and industry-specific factors in global value networks. Therefore, IB research often seeks to unveil the cause-and-effect implications (or, treatment effects) of critical events that shift these factors at the country, industry, or firm level for firms' global strategies and international operations. Moreover, as the impacts of these critical events in the IB contexts often last over years and affect different firms at different time points, IB research commonly relies on panel data to capture such treatment effects. In this regard, DID design offers a superior empirical solution in line with such key features of IB research. Commensurately, ever more studies using DID design with panel data have been published over the past decade. For example, among the 131 empirical studies published in the *Journal of International Business Studies* between 2020 and 2022, a sizable portion (18 studies, 13.0%) applied DID design with panel data as at least part of their empirical methodologies.

Notably, recent research has begun to recognize that, in many circumstances, DID design with panel data could obtain biased casual estimates and thus ought to be applied with caution (e.g., Goodman-Bacon, 2021; Sun & Abraham, 2021). In light of such methodological progress, it is both pertinent and valuable for us to review the progress and potential issues of the application of DID design with panel data in the extant IB literature in the spirit of offering practical guidance to help enhance the rigorousness of future IB research.

DID design with panel data

In practice, the simplest DID design (or a "2 × 2" design) is often estimated as follows:

Jiatao Li and Han Jiang contributed equally to this study.

✉ Han Jiang
jianghan@cuhk.edu.cn

✉ Rongjian Yu
yurongjian@zjgsu.edu.cn

Jiatao Li
mnjtli@ust.hk

Jia Shen
jia.shen@utdallas.edu

Haoyuan Ding
ding.haoyuan@mail.shufe.edu.cn

[1] Center for Business Strategy and Innovation, School of Business and Management, Hong Kong University of Science and Technology, Hong Kong, China

[2] School of Management and Economics, Shenzhen Finance Institute, The Chinese University of Hong Kong, Shenzhen, China

[3] Jindal School of Management, University of Texas at Dallas, Richardson, USA

[4] College of Business, Shanghai University of Finance and Economics, Shanghai, China

[5] College of Business, Zhejiang Gongshang University, Hangzhou, China

Published online: 02 September 2024

$$y_{i,t} = \alpha + \beta_1 Treated_i + \beta_2 Post_t + \beta_3 Treated_i \times Post_t + \varepsilon_{i,t} \quad (1)$$

where $Treated_i$ indicates whether subject i belongs to the treated or control group, and $Post_t$ indicates whether an observation is from a time t before or after the treatment. The estimand is the averaged change in the outcome among the treated subjects before and after the treatment ($y_{i,t}\,(pre) - y_{i,t}\,(post)$), i.e., the average treatment effect on the treated (ATT). If the ATT is significantly different from the outcome change of the control subjects, we can infer a significant treatment effect.

Take the Sino–US trade conflict starting in 2018 as an example. One may reasonably wonder if this shock could drive the United States to replace China as source of imports with other countries geographically adjacent to China (e.g., Vietnam, Thailand, India, etc.). Such treatment effects of the trade war can be tested with a 2 × 2 DID design based on the US annual import data in 2017 and 2018. In this design, the annual imports of the US from each of China-adjacent countries serve as the treated group, and the imports from other countries (excluding China) as the control group. Should US imports from China-adjacent countries, on average, increase more in 2018 than those from other countries, we can infer a significant treatment effect of the trade war in terms of driving the US to relocate its imports from China to adjacent countries (rather than to the rest of the world).

Notably, the validity of this 2 × 2 DID design rests on a key assumption: The assignment of treatment is *random and exogenous* to any pre-existing difference between the treated group and the control group. In the above example, it is apparent that whether a country was adjacent to China (which determined whether it was in the treated or control group) was not related to the breakout of Sino–US trade war in 2018. As such, the assumption of random and exogenous treatment is met.

Compared with the basic 2 × 2 DID design, DID design with panel data allows the outcome of interest for a treated subject to be observed in multiple time periods before and after receiving the treatment, thus allowing unobserved subject- and time-specific heterogeneities to confound with the treatment effects. In this regard, researchers often perform DID estimations with panel data using two-way fixed effects (TWFE) regressions to control for subject- and time-fixed effects. If all treated subjects receive the treatment at the same time, the TWFE DID estimation manifests as follows:

$$y_{i,t} = \alpha + \beta_1 Treated_i \times Post_t + \gamma_i + \delta_t + \varepsilon_{i,t} \quad (2)$$

where γ_i indicates subject-fixed effects, and δ_t indicate time-fixed effects. Note that the main effects of $Treated_i$ and $Post_t$ are fully subsumed into the subject-fixed and the time-fixed effects. The ATT is calculated as the variance-weighted average change in the outcome over multiple periods.

Along with the fixed-time treatment above, DID design with panel data also allows different subjects to receive the treatment in different time periods (i.e., time-varying treatments). In that case, the treatment group in each time period consists of subject-time observations of those receiving the treatment in that period. In the meanwhile, the control group for that period may consist of three scenarios: (1) the observations for subjects that never received treatment throughout the entire observation window (i.e., *never-treated subjects*), (2) those for subjects that received the treatment after the given period (i.e., *to-be-treated subjects*), and (3) those for subjects that received the treatment prior to the given period (i.e., *already-treated subjects*). In such a manner, the ATT estimand of such time-varying treatment is the variance-weighted average of a series of time-specific 2 × 2 DIDs from each time period in the observation window. With the TWFE estimation, it is estimated as follows:

$$y_{i,t} = \alpha + \beta_1 Post_{i,t} + \gamma_i + \delta_t + \varepsilon_{i,t} \quad (3)$$

where γ_i is the subject-fixed effects, and δ_t is the time-fixed effects. In this model specification, the post-treatment dummy $Post_{i,t}$ is equivalent to the interaction between $Treated_i$ and $Post_t$ in a fixed-time treatment design, the main effects of which are subsumed by the two fixed effects.

We can also illustrate the above DID design with panel data using the example of the Sino–US trade war. By extending the observation window to cover US annual imports before 2017 and after 2018, we can build a panel dataset consisting of sourcing country-year observations to test the treatment effects of the trade war on the relocation of US imports. In this design, all China-adjacent countries serve as the treated group, and all other countries (excluding China) as the control group. 2018 and subsequent years serve as the post-treatment period. If US imports from China-adjacent countries increased more quickly than those from other countries after 2018 after controlling for country- and year-fixed effects, we can infer a significant Granger causality through which the trade war stimulated the growing time-series trend of US import relocation to China-adjacent countries.

It is also worth noting that, in practice, the Sino–US trade war did not affect all industries all at once in 2018 as an overarching policy shock. Instead, the US government gradually expanded the list of entities and products to enforce punitive tariffs, thus involving more industries and firms in the trade war over time. As such, the impacts of the trade war on US imports can be tested at the industry level in a time-varying manner, with different industries receiving the treatment at different time points. Such industry-level treatment effects can be tested with a panel dataset consisting of industry-sourcing country-year observations. In each year, the treatment group consists

of the annual imports of the US from each of the China-adjacent countries in each industry newly added to the punitive tariff list in this year. The control group includes annual imports from all non-China-adjacent countries in all industries, those from China-adjacent countries in industries that were not yet added to the punitive tariff list in the given year, and those from China-adjacent countries in industries that were already on the list prior to this given year. If US imports from China-adjacent countries, after controlling for industry-, country-, and year-fixed effects, increased more quickly over time than those from other countries after an industry was added to the punitive tariff list, we can infer the industry-level treatment effects of the trade war on the import relocation of the US.

Like the 2 ×2 DID design, the validity of the above Granger causalities inferred from TWFE DID estimations also hinges on the *randomness and exogeneity* of the treatment assignment. In DID design with panel data, this assumption requires the treated and control subjects share the comparable or parallel time-series trends in the outcome of interest prior to the occurrence of the treatment (i.e., the "*parallel trend assumption*"). To meet this assumption, whether and when a subject receives the treatment should be random and free of time-varying confounders, and it cannot be related to the subject's past outcome of interest. In the second sample design above, this assumption requires that whether and when an industry was added to the punitive tariff list was not determined by the US imports in this industry from China-adjacent countries prior to the trade war. If this is the case, we can observe no systematic deviation in the time-series patterns of industry-level imports of the US from China-adjacent countries and other countries before an industry was added to the list.

Moreover, since the treated subjects in DID design with panel data can have multiple post-treatment observations, this design also requires the stability of post-treatment trends. Such stability is hinged on two assumptions that rule out systematic differences among never-treated, not-yet-treated, and already-treated observations in the control group from different time periods. First, the treatment occurring in a given time is assumed to only affect the treated subjects' outcomes in that period without lingering impact afterward (or, carryover effect). This is termed the *static treatment assumption*. Second, since different subjects may be treated at different times, the treatment effects are assumed to remain constant over time no matter when a subject receives the treatment. This is termed the *constant treatment assumption*. Violating either of the post-treatment trend assumptions would confound the already-treated observations and contaminate the control group, thus biasing TWFE DID estimations with panel data. In such a manner, for the second sample design above to generate unbiased estimations, it requires the trade war to drive the US to relocate imports only in the year of its breakout in a given industry, and to affect all industries equally over time.

DID design with panel data in IB research: Progress and issues

Notably, recent econometric studies (e.g., Baker, Larcker, & Wang, 2022) brought to our attention that, in prior studies using DID design with panel data, the violations of the three assumptions above are rather common. In research contexts filled with major events that can affect broad and diverse subjects at multiple levels and over long-time spans (like the IB contexts), the violations could be even more pronounced (Liu et al. 2021). To investigate whether DID design with panel data has been used in the IB literature as effective as it should be, we conduct a comprehensive literature search to systematically comb through all IB studies published on major academic journals between 2012 and 2022. This search identifies 59 studies that examined IB-related topics using DID design with panel data as at least part of their empirical methodologies (hereafter, IB-DID studies). Online Appendix 1 highlights the method of this literature search and briefly summarizes the 59 studies.

Among the 59 studies, 27 (45.76%) were published in and after 2020, including 18 studies published in *JIBS* between 2020 and 2022. This indicates the fast-growing application of DID design in IB research in recent years. This trend is also witnessed by the organization and strategy research. For example, 11% (26 studies, including 3 IB-DID studies) of the 232 empirical studies published in *SMJ* over the past 3 years adopted DID design with panel data. Considering this emerging trend, it is both timely and pertinent to grasp the state of the arts for this method to be applied in the IB literature and diagnose the issues thereof to enhance the rigorousness of future research.

A key difference across the DID design in the 59 IB-DID studies is the level of treatments. Specifically, 21 studies (35.6%) used a key event that occurred on the firm level as the treatment. Most of those firm-level treatments were based on a firm's first internationalization (its first overseas investment, the beginning to export, cross-listing in U.S., directors' first foreign experience, etc.) or a major change in the firm's ownership structure (e.g., an IJV becoming a wholly owned subsidiary, a local firm being acquired by MNEs, a government acquisition, an IPO, etc.). Notably, as those firm-specific events inherently occurred at different times for different firms, all 21 studies based on firm-level treatments adopted the DID design with time-varying treatments (as in Eq. 3 above).

The remaining 38 studies (64.4%) adopted treatments on the country level (32 studies), the regional level (two

studies), or the industry level (four studies). Those country-level treatments were based on major socioeconomic events that occurred on the national or supranational level (e.g., the 1998 Asian financial crisis, a natural disaster, elections, etc.), or the enactment of critical laws or policies in a given country. At regional level, both studies adopted treatments based on major political changes. Lastly, all four industry-level studies focused on tariff reductions in a given industry as the treatments.

Although these IB-DID studies offer valuable insights about firms' global strategies and international operations, our review also reveals alarming issues in their DID estimations with panel data in these studies, especially in terms of potential violations of the three critical DID assumptions.

Potential violation of treatment randomness and exogeneity

The first issue revealed by our review is that the treatments adopted in a sizable portion of the 59 studies may not be random or exogenous by nature. This issue was particularly prominent in IB-DID studies using firm-level treatments based on representative events in a firm's history (e.g., internationalization, ownership changes, etc.). These firm-specific events are inherently correlated with the treated firms' strategies and performance, which also affect the firms' outcomes of interest both before and after the treatments. The assignment of these treatments can thus be endogenous, such that whether a firm received the treatments is determined by the historical status of its outcomes of interest (i.e., the "feedback effect"). For example, several studies used firms' first overseas investment as a treatment, which may be endogenous in a firm's strategies and performance. Likewise, another treatment, local firms being acquired by foreign MNEs, can also be endogenous in the local firms' performance and quality. In both cases, the treated firms can be systematically different from control firms in terms of their operations, and outcomes, thus violating the randomness of treatments.

On a related note, 12 of the 21 studies using firm-level treatments (57.1%) did not report any form of parallel trend test at all. Moreover, only 30 studies (50.8%) reported tests or illustrations demonstrating parallel trends between the treated and control groups. Such lack of robust tests for parallel trend further exacerbates the above concern about non-random and endogenous treatments.

Potential violations of the static treatment assumption

Our review also shows that the treatments adopted in many of the 59 IB-DID studies can be subjected to carryover effects over time. These treatments may not only have contemporaneous impacts in the time period when the treated subjects received the treatments but can also continue to affect these subjects' outcomes in future periods (i.e., already-treated observations). Such carryover effects were particularly prevalent in studies adopting grand events like 9/11 or China's WTO entry as the treatments. The impacts of those grand events often take multiple years to be incorporated by firms. Moreover, there can often be a series of consequent changes and shocks following those events. As such, those treatments not only have lingering impacts, but can also vary over time in terms of range and magnitude. In the meanwhile, these grand events can often affect various facets of treated subjects' operations and performance, which may interact with the outcome of interest over time. Such heterogeneous treatment effects would bias the estimand of TWFE DID design with panel data (Liu et al. 2021).

Potential violations of the constant treatment assumption

It also comes to our attention that some of the IB-DID studies used treatments that may not necessarily have constant influences across different subjects and/or time periods. This issue was particularly prominent in studies using samples from multiple countries. For example, several studies constructed DID design based on the enactment of M&A laws in a given country. Such laws or regulations often have different terms and stipulations in different countries and may thus have different implications for firms in those countries. Likewise, another treatment widely adopted in prior IB-DID studies, tariff reductions in the U.S., can often vary across industries in terms of magnitude, length, and range, thus leading to the heterogeneous treatment effects of such policy shocks on firms in different industries. In both cases, the treatment effects are characterized by heterogeneities across treated subjects, which, as discussed, can contaminate the already-treated observations and thus bias the DID estimations.

Notably, only 22 (37.3%) of the 59 IB-DID studies strictly followed the standard model specification of TWFE DID estimation and accounted for both subject- and time-fixed effects, including nine studies using fixed-time treatments and 13 using time-varying treatments. Nine of the 59 studies (15.3%) did not control for any fixed effect at all in their DID analyses. Twenty-eight (47.5%, five with fixed-time treatments and 23 with time-varying treatments) controlled for either subject- or time-fixed effects but not both. As both the carryover and heterogeneities in the treatment effects are partially nested in subject- and time-fixed effects, failing to appropriately control for both

may exacerbate the bias caused by violation of the static treatment or the constant treatment assumption.

Illustrations on potential bias of DID design with panel data

Following prior studies (e.g., Baker et al., 2022; Bertrand, Duflo, & Mullainathan, 2004), we perform two Monte Carlo simulations of TWFE DID estimations using hypothetical treatments that violate the three key assumptions. Note that these simulations are not designed to replicate prior studies, but to illustrate the estimation biases that could be caused by the above issues in the DID design with panel data we detect in prior IB-DID studies. In line with this mission, without further specification, the parameters we adopt in the simulations hereafter are set in the spirit of increasing the clarity of illustrations. In both illustrations, the panel datasets are generated using data on U.S. public firms extracted from *Compustat*. The codes used to perform the simulations are available upon request.

Illustration 1: Import tariff reduction and R&D intensity

The first illustration focuses on an industry-level treatment used in several IB-DID studies, i.e., tariff reductions. We simulate the impacts of a tariff reduction on R&D spending by firms in an industry. Echoing the insights and findings of prior IB studies, we posit that the tariff reductions decrease treated firms' R&D intensity in our simulations. That is, the intensified foreign competition brought about by tariff cut can not only undermine domestic firms' sales and profitability, but also drive them to focus on the immediate competitive threats on the market and thus become more short-term oriented in their strategic decisions. As such, domestic firms often cut back the resources they deploy to innovation in response to tariff reduction, thus reducing R&D activities.

In practice, a tariff reduction is unlikely to be directly shaped by the operations and decisions of any single firm, thus qualifying as a random and exogenous treatment. However, as it often takes time for foreign competitors to gradually enter the domestic market after the tariff barriers are lowered, the influences of tariff reductions in an industry, as noted by prior studies, take multiple years to manifest. Moreover, it is also noted that different industries are not equally susceptible to the impacts of tariff reductions. As such, the treatment effects of tariff reductions can be inherently dynamic and non-constant across treated firms and time periods. In such a manner, tariff reductions offer a realistic IB setting to illustrate the estimation biases that can be caused by the violations of the static and constant treatment assumptions even when the treatment is random and exogenous.

Using the annual records of all U.S. publicly listed firms from 1979 to 2019 from *Compustat*, we construct an unbalanced panel dataset including 102,835 firm-year observations. The outcome of interest, R&D intensity (RDI for short), is measured as a ratio of a firm's R&D spending over its total assets. We decompose RDI into a firm-fixed effect, a year-fixed effect, and a residual. The firm-year observations are sorted into 90 categories based on a firm's two-digit SIC code. We randomly designate an industry as experiencing a tariff reduction in a given year (or not) to ensure that the treatments of tariff reductions are randomized and exogenous to pre-existing heterogeneities. Based on this setup, we compose six hypothetical treatment effects of tariff reductions, each of which illustrate the bias caused by heterogeneous treatment effect, carryover effect, and the combination thereof. Using each hypothetical treatment, we use Monte Carlo simulations to create 500 simulated panel datasets. We use TWFE DID regressions to estimate the ATT of tariff reductions on treated firms' RDI for each simulated dataset, with tariff reductions measured as staggered shock dummies set to 1 for all treated and already-treated firm-year observations in a given year, and 0 otherwise.

In the first two simulations, we manipulate this treatment of tariff reductions by designating all firms in 45 randomly selected industries as experiencing a tariff reduction in 1999. In Simulation 1-1, the treatment effect of the tariff reduction is assumed to be a static shock that decreases the RDI of all treated firms by 50% of the standard deviation of RDI (σ_{RDI}) in 1999. This fixed-time treatment meets all three assumptions of DID with panel data: random and exogenous, static, and constant. In contrast, the fixed-time tariff reduction in Simulation 1-2 is assumed to decrease all treated firms' RDI by a constant extent of 5% of σ_{RDI} for 1999 and each year afterward. The second treatment thus violates the static treatment assumption, but not the constant treatment assumption.

We then conduct four simulations in which the tariff reductions are set to take place over three time periods—1989, 1999, and 2009. For each period, we designate the firms in 30 random industries as experiencing a tariff reduction. In Simulation 1-3, we assume the tariff reduction to decrease the RDI of all treated firms by a constant extent of 50% of σ_{RDI} in the shock year but not afterward. This treatment thus meets all three assumptions of DID design with panel data.

The treatment effects of tariff reductions in Simulation 1-4 are also assumed to only decrease a treated firm's RDI in the year it experienced the reduction. However, such static treatment effects are set as varying across the three treatment

periods at the level of 50% of σ_{RDI} for 1989, 30% of σ_{RDI} for 1999, and 10% of σ_{RDI} for 2009. As such, it violates the constant treatment assumption.

The treatment effects in Simulation 1-5 are assumed to be constant across industries treated at different time periods. However, such constant treatments are set as continuously decreasing the RDI of treated firms every year after the tariff reductions at the level of 3% of σ_{RDI}. In this regard, it has a significant carryover effect and thus violates the static treatment assumption.

Lastly, the treatment effect in Simulation 1-6 is assumed to both change over the three treatment periods and have significant carryover effect. For the firms receiving the treatment in 1989, their RDI is assumed to increase by 5% of σ_{RDI} in each year afterward. The change is instead taken as 3% of σ_{RDI} for firms treated in 1999, and 1% of σ_{RDI} for those treated in 2009. The last hypothetical treatment thus violates both the assumptions of constant and static treatment effect.

Figures 1 and 2 depict the distributions of the 500 estimated ATTs using TWFE DID regressions for the six hypothetical treatments, respectively. In each simulation, the "true" ATT can be calculated as the *observation-average ATT*, which is the equally weighted average of the ATTs across all treated firm-year observations, and the *firm-average ATT*, which first calculates each treated firm's equally weighted average ATT, and average those ATTs across all treated firms. Both approaches can be deemed unbiased in practice. We depict the two "true" ATTs in each figure with dotted lines.

Figure 1 presents the estimations for Simulations 1-1 and 1-2. As depicted in Fig. 1a, the TWFE DID regressions obtain unbiased estimations in Simulation 1-1, which adopts the fixed-time treatment that is both static and constant. The 500 estimated ATTs average to − 0.176, which equals both the observation-average ATT and the firm-average ATT. In contrast, Fig. 1b shows that the estimated ATTs are seriously biased in Simulation 1-2, which adopts the fixed-time treatment that is constant but not static. The mean of the 500 estimated ATTs from TWFE DID regressions is − 0.133, with a standard deviation (SD) of 0.005. This distribution significantly deviates from the observation-average ATT (− 0.060) and the firm-average ATT (− 0.058), indicating that the estimations of TWFE DID regressions are substantially biased by the presence of carryover effects.

Figure 2 presents the estimations for Simulations 1-3 through 1-6, which all adopt the time-varying hypothetical

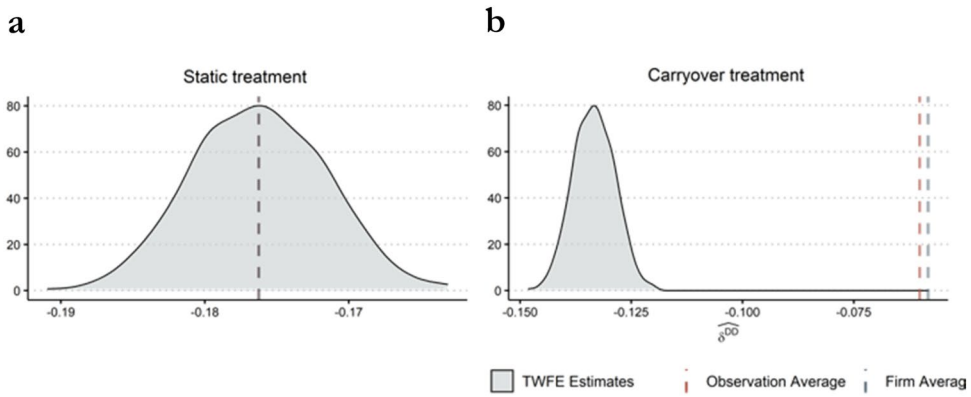

Fig. 1 Import tariff reduction and R&D intensity: fixed-time treatment

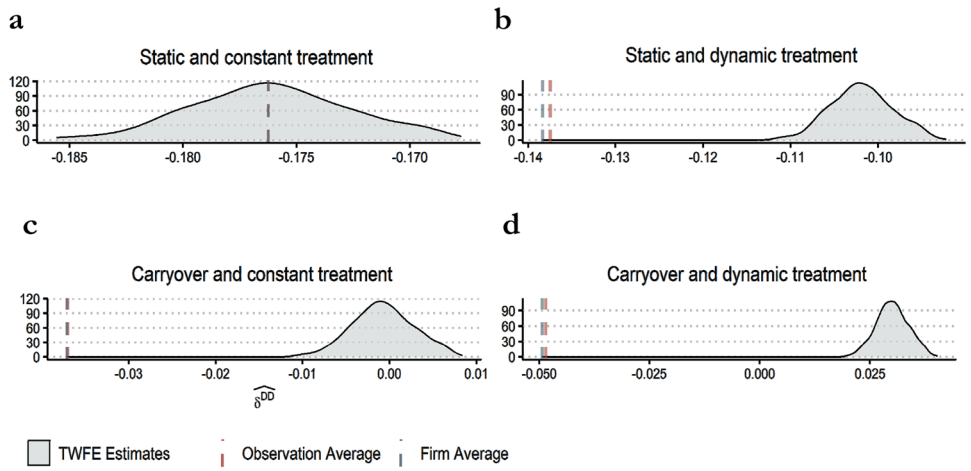

Fig. 2 Import tariff reduction and R&D intensity: time-varying treatment

treatments of tariff reductions. As depicted in Fig. 2a, the estimated ATTs of the 500 TWFE DID regressions average to − 0.176, which equals both the observation- and firm-average ATTs. This result indicates that TWFE DID regression can obtain unbiased estimations for the time-varying tariff treatment that meets all three assumptions for DID design with panel data.

In contrast, the estimated ATTs of the 500 TWFE DID regressions are severely biased in Simulation 1-4, in which the treatments are random, exogenous, static, but not constant across all treated firms. As depicted in Fig. 2b, the estimated ATTs average to − 0.102 (SD = 0.004), which significantly deviates from both the observation-average ATT (− 0.138) and the firm-average ATT (− 0.137). This result indicates that TWFE DID regressions will obtain biased estimations even when the time-varying treatments only violate the assumption of constant treatment effects.

Likewise, as shown in Fig. 2c, TWFE DID regressions fail to obtain unbiased estimations in Simulation 1-5, in which the treatments are random, exogenous, and constant, but have carryover effects over time. The estimation bias in Simulation 1-5 is even more severe than 1-4: In contrary to the significant and negative "true" ATT set in Simulation 1-5 (observation-average ATT = − 0.037; firm-average ATT = − 0.037), the mean of the 500 estimated ATTs is around − 0.001 (SD = 0.004) and statistically insignificant. This result indicates that compared with the violation of the constant treatment assumption, the violation of the static treatment assumption can be more problematic.

Lastly, Fig. 2d shows that in Simulation 1-6, in which both constant and static treatment assumptions are violated, the estimated ATTs are again significantly biased. In particular, TWFE DID regressions estimate significant and positive ATTs, with the 500 estimated ATTs averaging to 0.031 (SD = 0.004). However, as stipulated by the simulation setting, the observation-average ATT (− 0.049) and the firm-average ATT (− 0.048) are both significant and negative.

In sum, Illustration 1 cautions that DID design with panel data will obtain biased estimations when the treatments are not static or constant. Even though those treatments can be random and exogenous, using TWFE DID regressions to estimate their treatment effects is still problematic.

Illustration 2: Firms' internationalization and performance

The second illustration uses a firm-level treatment adopted in several IB-DID studies, i.e., the beginning of a firm's internationalization. We simulate the impacts of this treatment on firms' financial performance (ROA). The cause-and-effect link between internationalization and firm performance remains a fundamental debate in the extant IB literature. Using DID design with panel data, recent studies recognize that international operations not only expand the firms' accessible markets and resources, but also allow them to learn from foreign stakeholders to improve their operations and productivity, thus enhancing the focal firms' performance. Echoing such findings, we propose internationalization, as a treatment in our simulations, to improve firm performance.

However, a firm's internationalization is a self-selected decision inherently endogenous to its past strategies and operations. Such decision can be systematically correlated with the firm's past financial performance: Superior performance from the past can encourage a firm to engage in international expansion and affect the likelihood for the firm to receive the treatment (i.e., the "feedback effect"), thus violating the assumption of random and exogenous treatment. On top of that, it also stands to reason that the impacts of internationalization, as a fundamental strategic initiative, often manifest over a rather long period (non-static) and might vary across different firms (non-constant). Taken together, the treatment effects of internationalization on firms' financial performance offer a pertinent and realistic instance in the IB contexts to show the estimation biases that can be caused by potential violations of all three assumptions of DID design with panel data.

We construct a firm-year panel dataset using *Compustat* data from 2010 to 2019. The ROA of each firm-year observation is decomposed into a firm-fixed effect, a year-fixed effect, and a residual. The treatment assignment for internationalization is manipulated as follows: For each firm-year observation in year t, we first calculate the firm's ROA in year $t - 1$, along with the mean and SD of the lagged ROAs. The observations in year t are then grouped up based on sample firms' ROA in year $t - 1$, with each group of firms assigned a different likelihood of internationalization. Firms with a lagged ROA two SDs above the mean or higher are assigned a 90% likelihood of internationalization in year t. Such likelihood is set as 70% for firms with lagged ROA between one and two SDs above the mean, 50% for firms with a lagged ROA within one SD of the mean, 30% for firms with lagged ROA between one and two SDs below the mean, and 10% for all the other firms. The assignment of this hypothetical shock is non-random and endogenous, as the likelihood for a firm to embark on internationalization in a given year is set to be systematically correlated with its previous ROA.

Based on this setup, we compose four hypothetical treatments of internationalization, each of which illustrates the estimation biases caused by non-random and endogenous treatments, along with heterogeneous treatment effect, carryover effect, and the combination thereof. Like Illustration 1, we create 500 simulated panel datasets

for each hypothetical treatment and estimate the ATT thereof with TWFE DID regressions. Internationalization is measured as a staggered shock dummy valued as 1 for all treated and already-treated firm-year observations in a given year, and 0 otherwise.

The treatment effect of internationalization in Simulation 2-1 is assumed to increase the ROA of all treated firms by a constant level of 50% of the SD of ROA of all firms (σ_{ROA}) in the year in which they started internationalization (i.e., treatment year). Also, the simulated treatment is defined as statically affecting only the treatment year with no carryover effect afterward.

In contrary to the constant and static treatment above, Simulation 2-2 sets the treatment effects of internationalization to be static but varying over time in terms of magnitude. Specifically, for treated firms engaging in internationalization in 2011, we designate their ROA to increase by 90% of σ_{ROA}. Such effect is assumed to grow weaker by 10% of σ_{ROA} for firms treated in every year afterward (i.e., 80% of σ_{ROA} for firms treated in 2012, 70% of σ_{ROA} in 2013, and so forth). This setting echoes the non-random and endogenous nature of the hypothetical treatment assignment. That is, as internationalization is designated as a self-selected decision based on firms' performance concerns, it would stand to reason that firms' decision of later internationalization might be driven by the lesser benefits they could derive from such strategic initiative than early treated peers.

The treatment effect in Simulation 2-3 is designated as constant across all treated firms in terms of increasing their ROA by 5% of σ_{ROA} in the treatment year. On top of that, upon receiving the treatment, each already-internationalized firm is assumed to experience an increase in their ROA at the level of 5% of σ_{ROA} every year after the treatment year. With this setting, internationalization is designated to be constant but characterized with substantial carryover effects. Such non-static treatment effects also reflect the reality that it can take multiple years for firms to fully incorporate the benefits derived from their international expansion into their operations and performance.

Lastly, the treatment in Simulation 2-4 is defined as both non-static and non-constant. We designate the ROA of firms treated in 2011 to increase by 20% of σ_{ROA} upon being treated. Such treatment effect is set to grow weaker by 2% of σ_{ROA} for firms treated in every year afterward (18% of σ_{ROA} for firms treated in 2012, 16% of σ_{ROA} in 2013, and so forth). We designate this regressive treatment effect to be applied every year after a firm received the original treatment. By doing so, this hypothetical treatment in Simulation 2-4 violates all three DID assumptions.

Figure 3 depicts the distributions of the estimated ATTs of TWFE DID regressions based on the 500 Monte Carlo simulated datasets for each of the four hypothetical treatments. Again, the observation- and firm-averaged ATTs are plotted with dotted lines in each figure. As shown in Fig. 3a, in Simulation 3-1, the 500 estimated ATTs average to 0.173 (SD = 0.006), which largely deviates (2.8 SDs away) from both the observation-average ATT (0.190) and the firm-average ATT (0.190). Such significant bias indicates that TWFE DID regressions are unlikely to obtain unbiased estimations for an endogenous treatment, even when this treatment is static and constant.

Figure 3b through 3d further show that the biases caused by endogenous treatments will be exacerbated by the violations of the constant and static treatment assumptions. As shown in Fig. 3b, in Simulation 2-2 in which the endogenous treatment is set to be static but not constant, the estimated ATTs from TWFE DID regressions are subjected to more pronounced biases. That is, the 500 estimated ATTs average to 0.214 (SD = 0.006). The observation-average ATT

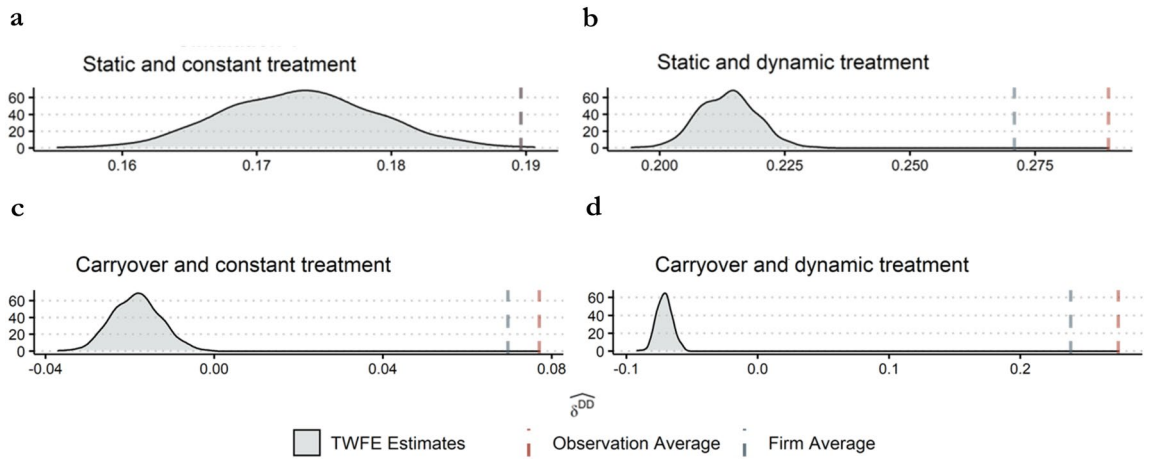

Fig. 3 Firms' internationalization and financial performance

(0.283) and the firm-average ATT (0.273) are both 10 SDs aways from the mean of the 500 estimated ATTs.

Moreover, in Simulation 2-3 in which the endogenous treatment is set to be constant but not static, the estimation biases of DID design with panel data are even more striking. As shown in Fig. 3c, TWFE DID regressions estimate the ATT of internationalization to be significant and negative (mean = − 0.019, SD = 0.006). However, in our setting, the observation-average ATT (0.062) and the firm-average ATT (0.059) are both positive. Such biases echo the findings of recent research (e.g., Baker et al., 2022), which note that DID design with panel data could obtain ATT estimations with opposite signs of the true ATTs with the presence of carryover effects in the treatment.

Lastly, Fig. 3d shows that in Simulation 2-4, in which the treatment is set to violate all three DID assumptions, the ATT estimations of TWFE DID regressions are again significant and of opposite sign. The estimated ATTs are significant and negative (mean = − 0.108, SD = 0.006), while the true ATTs are set to be positive (observation-average ATT = 0.219; firm-average ATT = 0.20).

In sum, Illustration 2 shows that when the treatment is endogenous, DID design with panel data will always obtain biased estimations. Such biases can be further intensified by the violation of the other two assumptions of DID designs, especially the carryover effects in the treatments.

Practical recommendations

Research design: Choose IB treatments carefully

Firm-level treatments

As we highlighted above, many IB-DID studies adopted major events in firm histories as the treatments in DID design with panel data. However, this practice should be treated with great caution. Because these firm-level events are commonly endogenous to a firm's historical paths of operations and performance, they are often correlated with various firm-specific factors (especially those steering the firms' IB practices) and thus subjected to potential feedback effects with the outcomes of interest over time. As a result, DID design based on these events often violates the assumption of random and exogenous treatment. Moreover, firm-level treatments can vary across firms and time periods, thus violating the constant treatment assumption. In addition, because a major event in a firm's history often has lingering impact on numerous facets of the firm's operations and outcomes, it often has significant carryover effects and therefore violates the static treatment assumption. As demonstrated in Illustration 2, such potential issues of firm-level treatments can render the ATT estimations of DID design with panel data fundamentally biased.

In fact, in light of the severity of the above issues, DID design with panel data might not be the optimal empirical design for firm-level treatments. Instead, other identification strategies (e.g., two-stage regression with instruments, regression discontinuity design (RDD), etc.) may allow scholars to obtain more rigorous estimations for the cause-and-effect links of firm-level events in the IB contexts. Take another popular treatment in prior studies, i.e., local firms being acquired by MNEs, as an instance. Instead of using these events as DID treatments, scholars may only examine the "treated" firms acquired by MNEs, focusing on the strategic implications of key features of such cross-border acquisitions. For example, scholars may examine how the acquired firms' performance is affected by their foreign ownership, which can be instrumented by the industry average thereof (excluding the focal firms). Scholars may also use 50% of foreign ownership as the discontinuity to perform RDD for the impacts of foreign ownership on local firms' post-acquisition performance.

Industry- and country-level treatments

Compared with firm-level treatments, major events in industrial institutions or socioeconomic contexts at the national or supranational level in the IB contexts may offer more suitable settings in which DID design with panel data can be appropriately applied, as these industry- and country-level events are generally exogenous to any single firm's decisions or performance. That said, scholars still need to remain cautious about the randomness and exogeneity of such industry- and country-level treatments, especially those based on major changes in business policies or institutions. As highlighted by prior studies, governments and policy makers often make such changes based on certain systematic patterns in the operations or outcomes of firms in a particular industry or region. For example, governments commonly adjust the tariffs based on the historical import records and domestic firms' performance in an industry. Likewise, the enactment of new cross-border M&A regulations in a country can be driven by foreign acquirers' activities in this country from the past. In these cases, there might exist heterogeneities between the treated and control groups that may render the pre-treatment trends unparallel. In such a manner, it is critical for scholars to carefully choose the outcomes of interest in their IB-DID design based on industry- or country-level treatments in the spirit of avoiding potential feedback effects.

What calls for greater caution when using these treatments is the potential violation of the static treatment assumption. As we highlighted above, as these macro events commonly have long-lasting influence on all firms in the

treated industries, regions, or countries, their treatment effects are intrinsically subjected to carryover effects. Such carryover effects can be particularly prominent for grand national and supranational shocks in the IB contexts. For example, it is widely noted that the influence of a country's entry into and exit from major supranational organizations (e.g., China's WTO entry and Brexit) would generally affect the firms in this country for decades. Likewise, recent studies recognize that the global supply chains and value networks nowadays are still incorporating the impacts of the Sino–US trade war starting in 2018. As shown above, using non-static treatments with prominent carryover effects in DID design with panel data could be particularly problematic.

In addition, these industry- or country-level treatments may also risk violating the constant treatment assumption, as they often unfold distinctly in different industries, regions, or economies in terms of manifestations and magnitudes. Such potential violation can be particularly prominent in IB studies using samples from multiple countries. For example, it is noted that the impacts of a country joining the European Union on its firms' foreign market entry largely vary across countries joining EU at different times. Likewise, the recent global pandemic would affect the global supply chains of firms from different countries and different industries in majorly different ways. These issues not only bias the DID estimand but may also loom in a study's theorization as alternative explanations. In this regard, scholars need to closely delineate the sources of carryover effects and heterogeneities of their non-firm-specific treatments, both in theoretical development and in empirical estimations.

Analytical strategies: Adopting rigorous protocols

Perform appropriate parallel trend tests

The prominence and prevalence of the above issues urge scholars to diligently adopt rigorous protocols when using DID design with panel data in the IB contexts. First and foremost, it is critical for scholars to conduct appropriate parallel trend tests to gauge the randomness and exogeneity of the treatments they adopt. In such tests, the treated group and the control group must show no systematic difference in the outcome of interest prior to receiving the treatment. Notably, parallel trend tests are needed in both fixed-time and time-varying treatment designs. For fixed-time treatment designs, the common practice is to depict and compare the pre-treatment trend for both the treated and the control groups. For time-varying designs, Bertrand et al. (2004) propose the following parallel trend test:

$$y_{i,t} = \alpha + \beta_1 Post_{i,t-2} + \beta_2 Post_{i,t-1} + \beta_3 Post_{i,t} + \beta_4 Post_{i,t+1} + \beta_5 Post_{i,t+2} + \gamma_i + \delta_t + \varepsilon_{i,t}$$

where $Post_{i,t-n}$ is the pre-trend dummy valued as 1 for a treated subject n years prior to its first receiving the treatment and 0 otherwise. $Post_{i,t+n}$ is the post-trend dummy valued as 1 for a treated subject n years after its receiving the treatment and 0 otherwise. Note that the number of such pre- and post-trend dummies can be adjusted based on the empirical setting of a given study. These pre-trend dummies contrast the n-year pre-treatment trend between the treated group and the control group. Should these dummies have no significant impact on the outcome of interest, it indicates that the treated sample's pretreatment trend of the outcome of interest is not significantly different from that of the control group. The random and exogenous treatment assumption is therefore satisfied.

Thoroughly control for fixed effects

Another protocol that was not always followed in prior IB-DID studies, as recognized by our review, is to account for fixed effects at all levels (especially subject- and time-fixed effects). These fixed effects at least partially subsume the carryover effects and heterogeneities in treatments, which are enrooted in the unique and non-varying heterogeneities of firms, years, industries, or countries. As such, failing to control all relevant fixed effects can exacerbate the violations of the constant and static treatment assumptions. This is particularly crucial for time-varying designs, in which the post-treatment dummy takes the place of the DID interactor.

In fact, it is widely noted that firms' international operations and global strategies are jointly determined by firm-, industry-, and location-specific factors, as well as temporal dynamics. In line with the nature of IB practices, IB scholars may try to account for the interactive fixed effects as a more robust approach to capture the interplay across unobserved specificities at different levels. Doing so can help better control for the potential non-static and non-constant treatments in DID design with panel data. For example, it is recognized that global supply chain resilience is shaped by firms' industrial conditions, as well as the locational specificities in both their home countries and sourcing countries. In this regard, to examine the impacts of the global pandemic on the survival of global supply chains of MNEs, scholars can control for the interactive fixed effects across industries, home countries, and sourcing countries, along with the regular firm- and year-fixed effects.

Use matched samples

A common practice in prior studies using DID design with panel data is to construct matched control samples in the spirit of alleviating potentially non-random treatment assignments. This approach is also adopted by several IB-DID studies in our review. The rationale of this approach

is to match each treated subject with one or several control subjects that share comparable likelihood of being treated or similar key features with the treated subject (e.g., firm fundamentals, geographic locations, industrial affiliations, etc.). The matched samples are often constructed using propensity score matching (PSM) or coarsened exact matching (CEM). This approach can indeed help mitigate the estimation biases caused by the potential violation of random and exogenous treatment assumption, given that the matching criteria can accurately capture the major causes for non-random and endogenous treatment assignments. For firm-level treatments, the historical records of the outcomes of interest should always be incorporated in the matching criteria to control for feedback effects. For example, in Illustration 2, each treated firm in each period should be matched with a group of never- or not-yet internationalized firms with similar historical ROA, which was the key cause for the non-random treatment assignment in the setting.

However, due to the complexities of the IB contexts and firms' global strategies and cross-border operations, it is hard for researchers to thoroughly identify and account for all sources of endogeneities in treatment assignments in the matching process in practice, especially for firm-level treatments. More importantly, using matched sample as the control group cannot address the biases caused by non-constant or non-static treatments, which only affect the treated subjects. In such a manner, while the matched sample approach offers a valuable protocol to improve the rigorousness of DID design with panel data, IB scholars should trust its effectiveness with caution. Scholars need to diligently identify and adopt appropriate criteria to construct the matched samples.

Consider alternative estimators: CSDID and stacked regression estimators

Along with the above protocols, scholars have also recently developed alternative estimators to remediate the violations of DID assumptions and enhance the robustness of DID design with panel data. In particular, Callaway and Sant'Anna (2021) developed an estimator that is later known as the "CSDID estimator". It was originally designed to cope with heterogeneities of time-varying treatments in DID design with panel data. The CSDID estimator first estimates a "cohort treatment effect" for each cohort of subjects treated in the same time period. That boils down the subject-time observations into a series of time-specific cohorts, each of which provides a simple 2×2 DID design with clean controls consisting of never-treated or not-yet-treated observations. All cohort-time-specific treatment effects are then aggregated into a cohort-average ATT as the estimand of interest.

Another popular estimator is called "stacked DID" (Cengiz, Dube, Lindner, & Zipperer, 2019). This estimator first generates a period-specific panel data for each treatment period that includes both the observations treated in this period and the clean control observations that are untreated both before and shortly afterward. This excludes all not-yet-treated and already-treated observations from the period-specific panel dataset, thus constructing a clean fixed-time DID design for each treatment period. All the period-specific panel datasets are then stacked together into a new panel data to perform TWFE DID estimation. Notably, this TWFE DID estimation no longer controls for firm-fixed effects, but accounts for period-specific firm-fixed effects (or firm-by-period fixed effects) instead.

To gauge the rigorousness of these two alternative estimators when DID assumptions are violated, we replicate our two illustrations by using these two estimators to replace regular TWFE DID estimations. Figures 4 and 5 depict the replication results for the two illustrations, respectively (see also Online Appendix 2). As shown in Fig. 4, in Illustration 1, both estimators constantly obtain more rigorous estimations than TWFE DID regressions. For all six simulations, both the observation- and the firm-average ATT of the treatment effects of tariff reduction fall within one SD of the 500 estimated ATTs obtained with both CSDID and stacked DID. These results together show that when the treatment assignment is random and exogenous, both CSDID and stacked DID can effectively remediate the violations of the constant and static treatment assumptions.

However, when the treatment assignment is non-random and endogenous, CSDID may no longer be rigorous. As shown in Fig. 5, the "true" ATTs of the endogenous treatment of internationalization in Illustration 2 constantly fall beyond 8 SDs away from the average of the 500 estimated ATTs obtained with CSDID, indicating that CSDID is unlikely to obtain unbiased estimations whether the other two assumptions are violated or not. In contrast, for this endogenous treatment, stacked DID obtains much more robust estimations than both TWFE DID regressions and CSDID. Both the observation-average ATT and the firm-average ATT still constantly fall within 1 SD from the average of the 500 estimated ATTs obtained with the stacked DID estimator.

In light of the above findings, it can be valuable for future IB scholars to consider those alternative estimators (when applicable) in their DID design with panel data at least as robustness checks. Especially, for DID design based on firm-level treatments, using stacked DID estimator in place of regular TWFE DID regressions may effectively alleviate the potential estimation biases.

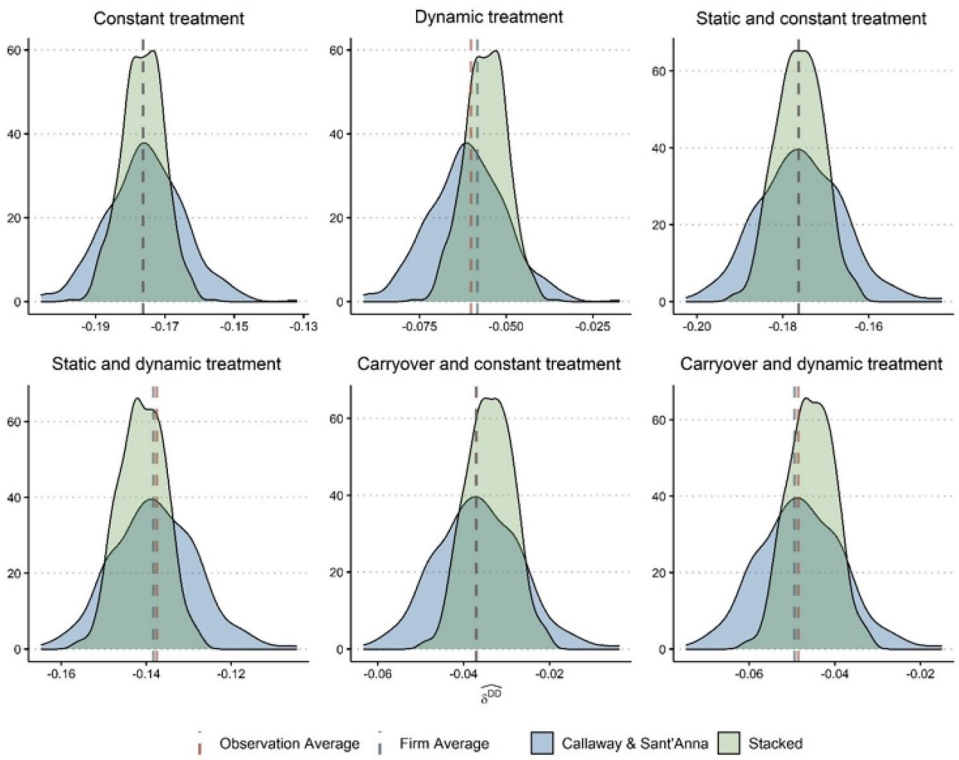

Fig. 4 Alternative estimators of tariff reduction and firm innovation

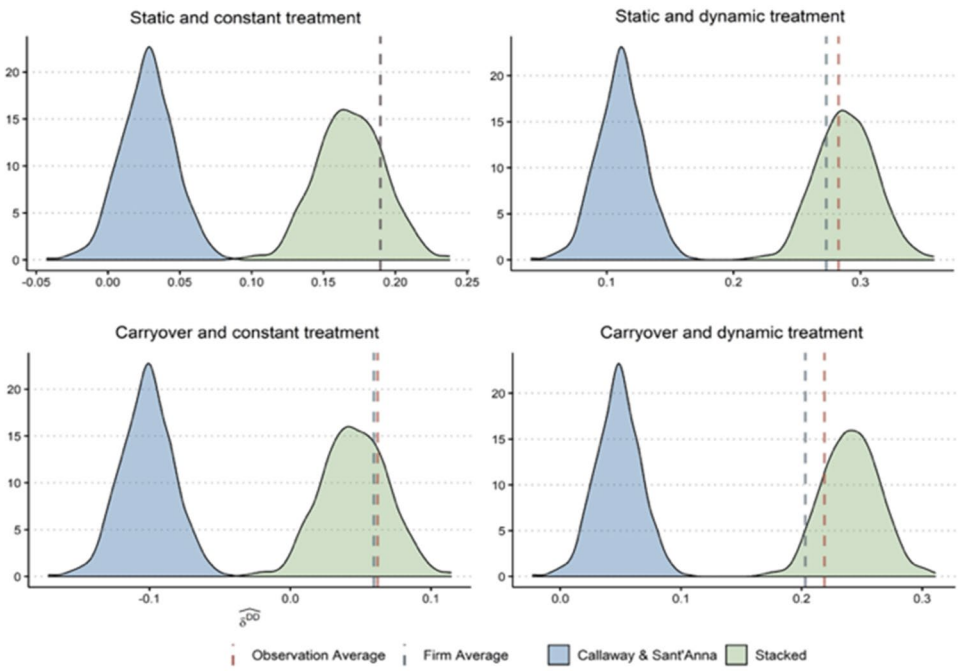

Fig. 5 Alternative estimators of firms' internationalization and performance

Conclusion

DID design with panel data has become increasingly popular in IB research over the past decade. However, our review recognizes that in the extant IB-DID studies, the assumptions of random and exogeneous treatment, static treatment, and constant treatment are often violated, which can cause drastic estimation biases. We offer a series of practical recommendations to help future IB researchers enhance the rigorousness of DID design with panel data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1057/s41267-024-00725-3.

Acknowledgements We thank the JIBS EIC, Prof. Rosalie Tung, and two anonymous reviewers for their insightful comments. This editorial was supported in part by the Research Grants Council of Hong Kong (HKUST# 16506622) and by the National Natural Science Foundation of China (No. 71973129).

References

Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics, 144*(2), 370–395.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics, 119*(1), 249–275.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230.

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *Quarterly Journal of Economics, 134*(3), 1405–1454.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277.

Liu, L., Wang, Y., & Xu, Y. (2021). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. arXiv preprint arXiv:2107.00856.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2), 175–199.

Jiatao Li is Chair Professor of Management, Lee Quo Wei Professor of Business, Lee Heng Fellow, Director of the Center for Business Strategy and Innovation, and Senior Fellow of the Institute for Advanced Study, HKUST. He is a Fellow of the AIB and an editor of JIBS. His research interests are in the areas of global strategy, innovation, entrepreneurship, corporate governance, and digital economy.

Han Jiang is Associate Professor and Presidential Fellow at the Chinese University of Hong Kong (Shenzhen). His research focuses on corporate governance, global strategies, entrepreneurship, and quantitative research methods. His work has been published in the *Academy of Management Journal*, *Strategic Management Journal*, *Production and Operations Management*, *Journal of Management*, *Journal of Business Venturing*, *Entrepreneurship Theory and Practice*, among others.

Jia Shen is a PhD candidate in the Department of Organization, Strategy, and International Management at Naveen Jindal School of Management, University of Texas at Dallas. Her research interests focus on corporate governance, global strategies, entrepreneurship, and innovation.

Haoyuan Ding is Professor and Associate Dean of the College of Business at SUFE. He serves as an associate editor of the *Journal of International Money and Finance*, *China Economic Review* and *China & World Economy*. His research interests lie in the fields of international trade, foreign direct investment and global supply chain management.

Rongjian Yu is Professor and Dean of the School of Business Administration, Director of the ChinaChain Research Center at Zhejiang Gongshang University, and Chair of the International Business Professional Committee in China. His research interests focus on global supply chains' innovation, digitalization, and restructuring.