Presenting and Interpreting Event-Study Betas in the H191 Thesis

This report translates the guidance from recent multi-period difference-in-differences and event-study papers into a concrete strategy for how you should **present and interpret the β coefficients** (and the associated R^2) in your thesis tables and figures. The focus is on:

- Callaway & Sant'Anna (2021, 2022)
- Sun & Abraham (2021)
- Borusyak, Jaravel & Spiess (2022)
- Roth (2024)

and how their insights apply to your **port×terminal×month** event-study design with staggered reforms, not-yet-treated controls, and heterogeneous treatment effects.

Throughout, I'll refer to your event-time index as **m** (with m = 0 the reform month, m < 0 leads, m > 0 lags), your outcome as **ln(LP)**, and your main summary parameter as the **average post effect over m ∈ [1, 12]**.

---

1. Conceptual background: what your βs are supposed to measure

All four papers start from the same problem: with staggered treatments and heterogeneous effects, **naïve TWFE event-study regressions do not return a simple "average treatment effect at horizon m"**. Instead they mix many cohort- and time-specific effects with non-transparent, sometimes negative weights.

Your thesis already avoids the worst of this by:

- Using **not-yet-treated (NYT) comparison sets** under each event clock.
- Estimating **event-time coefficients β(m)** relative to an omitted month m = −1.
- Then aggregating these β(m) into an average post-treatment effect over m ∈ [1, 12].

Conceptually, you want β(m) and their aggregates to be interpretable as:

> "Average causal effect of the reform on ln(LP) **h months after the reform**, for the relevant treated units, relative to the counterfactual evolution of those units had they not been treated."

The papers you cited provide three key pieces of guidance:

1. **Define an estimand first** (what population/time average do you want?), then choose an estimator that delivers it with transparent weights (Callaway–Sant'Anna; Borusyak–Jaravel–Spiess).
2. **Present the dynamic path in a way that makes the weighting clear** and separates pre-trends (leads) from post-treatment dynamics (lags) (Sun–Abraham; BJS).

3. **Be explicit about what you test with pre-trend F-tests and how you interpret them** (Roth, plus Sun–Abraham's discussion of differential trends).

Your tables should therefore:

- Make clear **what average over m** you are reporting (e.g., simple average of β(m) over m ∈ [1, 12]).
- Clarify **what is being averaged over units/cohorts** (terminals, ports, months), and whether there is a size weight (e.g., TEU-weighted vs unweighted).
- Maintain a tight link between the **event-study plot** (dynamic β(m)) and the **summary rows** ("Average post m ∈ [1, 12]", "Implied %ΔLP").

---

2. Lessons from Callaway & Sant'Anna: group-time ATTs and aggregation

2.1. Group-time ATTs as the building blocks

Callaway & Sant'Anna (2021) propose that in staggered DiD we should think in terms of **group-time average treatment effects**, ATT(g, t):

- **g** indexes a **cohort** (e.g., the month/terminal when the reform starts).
- **t** indexes **calendar time**.

Their main point: **all interesting causal parameters in staggered DiD can be written as averages of ATT(g, t) with transparent, non-negative weights.** For example:

- "Average effect over all treated months" ⇒ average of ATT(g, t) over (g, t) pairs where units are treated.
- Event-study profiles ⇒ average of ATT(g, t) over all (g, t) with the same **event-time h = t − g**.

Your β(m) should be thought of as empirical analogues of these **ATT(h)** parameters: average ATT(g, t) at event time m = t − g for the relevant set of groups.

2.2. Aggregation choices: simple vs. group-weighted vs. calendar-time weighted

The **did** package and the associated vignette (C&S 2021/2022) highlight several standard aggregation schemes for ATT(g, t):

- **Simple average ATT**: equal weights on all ATT(g, t) in the estimand set.
- **Dynamic (event-time) average ATT(h)**: equal weights on ATT(g, t) with the same horizon h.
- **Group-weighted averages**: weights proportional to group size.
- **Calendar-time averages**: weights by the share of treated units in each calendar time.

Applied to your thesis, this gives you a menu:

1. **Dynamic path β(m)**

2. For each event month m, you estimate a β(m) that is effectively an **ATT(m)** across terminals/ports that are at that event time.
3. This is what goes into your **event-study figures and Appendix dynamic tables**.
4. **Average post effect over m ∈ [1, 12]**
5. In Tables 1 and 5, you are reporting an **average over event-time horizons** in the first post year.
6. The cleanest choice (aligned with C&S) is:
    - **Simple equal-weight average over m** (each month 1,...,12 gets weight 1/12), and
    - Within each m, β(m) is itself an unweighted average across treated terminals at that horizon (or, if you prefer, TEU-weighted; see below).
7. **Weighting across terminals**
8. C&S emphasise that you should choose weights that match the **economic question**.
9. For you, two natural options:
    1. **Unweighted terminal average**: each treated terminal contributes equally → good if you care about the "average terminal" effect.
    2. **TEU-weighted** average: weight units by throughput → good if you care about the average **container handled** being processed more or less productively.

Recommendation for main text

- For the **main "Average post m ∈ [1, 12]" rows**, use a **simple average over β(m) for m = 1,...,12**, with **unweighted terminals** (your default). Make this clear in a note ("simple average over event months m = 1,...,12; each terminal weighted equally").
- In **robustness/appendix**, consider a **TEU-weighted alternative** ("Avg post (TEU-weighted)" row or separate table) to show that the results are not driven purely by small or large terminals.

This mirrors C&S's approach of having a **primary aggregation choice** and then checking robustness to alternative weighting schemes.

---

3. Lessons from Sun & Abraham: dynamic profiles and avoiding TWFE contamination

Sun & Abraham (2021) focus on the fact that **standard TWFE event-study regressions produce misleading dynamic treatment paths when effects are heterogeneous**. They show that:

- TWFE leads and lags are **linear combinations** of cohort-specific ATT(g, t) with **non-convex, sometimes negative weights**.
- Lead coefficients can be contaminated by post-treatment periods of early-treated groups, making pre-trend tests hard to interpret.

Your design avoids the main TWFE pathologies because:

- You adopt a **NYT comparison design** under each clock, rather than including already-treated units as controls.
- You work with **event-time indicators** relative to the port-specific reform date, with m = −1 omitted.

Sun & Abraham's contribution for you is more about **how to present the dynamic β(m)**:

1. **Separate leads and lags visually and in tables**
2. In figures, use different shading or a vertical line at m = 0; group leads and lags.
3. In tables, group pre-treatment rows (e.g. "(−4…−2) avg") separate from post-treatment rows.
4. **Summaries of post-treatment horizons**
5. They often summarise dynamics by reporting coefficients at selected horizons (e.g., m = 0, 1, 3, 6, 12) and sometimes an **average over a window**.
6. This supports your design choice of having:
    - **Selected event months** (0, 1, 3, 6, 12) in the Haifa privatization table, and
    - **Average post m ∈ [1, 12]** in Table 1 and Table 5.
7. **Pre-trend diagnostics**
8. Sun & Abraham emphasise that **non-significant leads do not prove no pre-trends**, but significant leads are strong red flags.
9. Hence, it is good practice to:
    - Report **both individual lead coefficients** (in plots/tables) and
    - A **joint F-test p-value for all leads** (your "Pre-trends: p(Leads F-test)" row).

How this maps to your tables

- **Main tables (Table 1 & 5)** should include:
- "Average post m ∈ [1, 12]" (simple average ATT over first year).
- "Pre-trends: p(Leads F-test)" to summarise evidence of differential pre-trends.
- **Dynamic tables (Appendix A)** and figures should:
- Show the full grid of β(m) with standard errors or confidence intervals.
- Include a row that averages key pre-treatment leads (e.g. (−4…−2) avg) to visually check approximate flatness.

This presentation is directly in line with Sun & Abraham's emphasis on **transparent dynamic profiles** and treating leads as genuine pre-trend diagnostics.

---

4. Lessons from Borusyak, Jaravel & Spiess (BJS): estimands, imputation, and summary measures

Borusyak, Jaravel & Spiess (2022) propose an **imputation-based event-study estimator** that explicitly reconstructs **untreated potential outcomes** for treated units and then averages treatment effects across cohorts and time.

Key points that matter for your thesis:

1. **Estimand clarity**
2. BJS insist that researchers should **start from a clearly defined estimand**:
    - For dynamics: ATT(h) for h in some horizon set.
    - For overall effect: an average ATT over some event-time and/or calendar-time window.
3. **Horizon-specific vs. window-average effects**
4. They emphasise that **single-horizon β(h)** are useful to understand timing, but often the economically relevant question is about **average effects over a relevant window** (e.g., first year, medium run, long run).
5. Hence, it is standard to:
    - Show the **full horizon profile in a figure**.
    - **Summarise** in text/tables with one or a few aggregate measures (e.g. avg over h ∈ [1, 4], [5, 8]).
6. **Weights and sample support**
7. BJS highlight the importance of **sample support at each horizon**: near the edges of the calendar window, fewer cohorts contribute to β(h), which can make estimates noisy.
8. This motivates reporting **N(h)** (number of comparisons per horizon) or at least noting that later horizons are estimated from fewer cohorts.

Application to your design

- You have chosen **m ∈ [1, 12]** as your primary post window, which is a clean and economically meaningful choice (“first year after reform”). This is exactly the kind of window BJS recommend summarising.
- You could supplement this with **window-sensitivity checks** in an appendix:
- Average post over m ∈ [1, 6] vs. m ∈ [7, 12], or [1, 24] if sample span allows.
- If you want, an Appendix table like: $\bar{\beta}_{[1,6]}$, $\bar{\beta}_{[1,12]}$, $\bar{\beta}_{[1,18]}$.
- In your **Appendix dynamic tables**, you already have an $N(m)$ column. That is directly in the spirit of BJS’s emphasis on support.

So your **main tables** should keep "Average post m ∈ [1, 12]" as the primary summary, while **Appendix tables** and possibly Table 5 can show a small number of alternative windows or weighted variants (TEU-weighted) to show robustness.

---

5. Lessons from Roth (2024): pre-trends and how not to over-interpret them

Roth's recent work (e.g., "Pretest with Caution") is about the **pitfalls of using pre-trend tests as specification diagnostics**.

Main takeaways relevant for you:

1. **Low power of pre-trend tests**
2. Failure to reject the null that all lead coefficients are zero **does not guarantee** the absence of economically meaningful pre-trends.
3. **Multiple testing / selection bias**
4. If researchers try many specifications and keep the one where pre-trend tests "pass", they induce **selection bias** and over-optimism in the reported post-treatment effects.
5. **Better practice**
6. Report **pre-trend coefficients and confidence intervals**, not just p-values.
7. Interpret pre-trend evidence **quantitatively** ("pre-treatment differences are within ±X% and statistically imprecise"), rather than as a simple pass/fail.
8. Avoid throwing away specifications purely because pre-trends are marginally significant; instead, **discuss the direction and size** of the pre-trends and whether they threaten your interpretation.

How this should shape your tables and text

- Keep your **"Pre-trends: p(Leads F-test)"** rows, but:
- In the notes, clarify that these tests are **informative but not definitive**; they are one diagnostic aligned with Roth's cautionary advice.
- In the text, when discussing results, say things like:
    - "Lead coefficients are small and statistically imprecise, suggesting no strong pre-trends, but the tests are underpowered given only 2 ports."
    - Or: "The leads show a moderate upward drift in LP even before the reform; this suggests caution in interpreting the post-reform increase as purely causal."
- In your **Appendix pre-trend table (S1)**, you already have structure to summarise:
- number of leads, p-values for lead F-tests, and placebo clocks.

- You can add a short note emphasising the Roth perspective on **interpretation**, not selection.

---

6. Concrete recommendations for your tables

Here is how I would implement all this in the specific tables you've drafted.

6.1. Table 1 – Main average post-reform effects

Rows to keep (core):

- **Average post m ∈ [1, 12]**
- Definition (in notes): simple average of event-time coefficients over m = 1,…,12, with equal weight on each month, and each treated terminal weighted equally.
- Interpretation in text: "Average effect of the reform on ln(LP) in the first year after treatment."
- **Implied %Δ LP**
- Definition: 100·(e^β − 1).
- Interpretation: point estimate of the **percentage change in LP** in the first post-year, relative to the counterfactual.
- **Pre-trends: p(Leads F-test)**
- Definition: p-value for H_0: all lead coefficients = 0.
- Interpretation: used as a descriptive diagnostic, not a binary spec test.
- **Observations**
- Total number of terminal×month observations used.
- **Within R^2**
- Definition (note): "Within R^2 from the underlying fixed-effects regression (variation within terminals over time)."

Rows to drop (as you already started doing):

- Terminal FE, Month FE: always present by design.
- Comparison set (always NYT under each clock).
- Port trends row (col headers already encode whether +PortTr is included).

Optional additional rows:

- If you implement TEU-weighted aggregation, add a row under each panel:
- "Avg post m ∈ [1, 12] (TEU-weighted)" with its own β(SE) and Implied %ΔLP row.
- This can be in the main table or in a robustness table, depending on space.

6.2. Table 2 – Haifa privatization dynamics

This is your "mini-dynamic" table for one specific clock (Haifa privatization). Given Sun & Abraham and BJS, the current structure is already good:

- **Panel A: Haifa Legacy under privatization clock**
- Average post m ∈ [1, 12] and Implied %Δ LP.
- Selected months: m = 0, 1, 3, 6, 12.
- Leads F-test: p (for pre-trends).
- **Panel B: SIPG (Bayport) placebo**
- Same structure, but you expect coefficients to be near zero.

Recommendations:

- In the table note, emphasise that this follows **Sun & Abraham-style dynamic presentation**, using the Haifa privatization as the event.
- Consider adding a **row for (−4…−2) avg** if you have enough leads; this compresses the lead evidence into a single number, similar to your Appendix dynamic tables.
- In the text: interpret the key βs as "short-run" (m = 0,1), "medium-run" (m = 6,12) and emphasise how they relate to the **average first-year β**.

6.3. Table 5 – Robustness of average post effects

Table 5 is your robustness table for **average post m ∈ [1, 12]** across different specifications.

For each panel (Haifa, Ashdod, Pooled):

- **Keep:**
- Entrant avg post m ∈ [1, 12] (β(SE) and implied %ΔLP).
- Legacy avg post m ∈ [1, 12] (β(SE) and implied %ΔLP).
- Observations and Within R^2.
- **Optional additions:**
- A second pair of rows for **TEU-weighted averages** (if you want to emphasise throughput weighting).
- One or two additional windows (e.g., "Avg post m ∈ [1, 6]" and "Avg post m ∈ [7, 12]") in an Appendix table rather than this main robustness table, to avoid clutter.

- **Notes:**

- Explicitly explain that “Base” and “+PortTr” correspond to the same specification as Table 1, and that “+Tr&Shocks”, “Balanced”, “Excl. 20–21”, “Excl. 23–24” are alternative sample/controls following BJS-style sensitivity to sample support and shocks.

6.4. Appendix dynamic and diagnostic tables

- **Appendix A (Full dynamic grids)**
- Keep the full β(m) table with N(m).

- Optionally add an **“average pre” row** (e.g. average over m ∈ [−4, −2]) and **“average post (1, 12)” row** for quick reading.

- **Appendix B (Pre-trend and placebo diagnostics)**

- Maintain the summary of leads, lead F-tests, and placebo clocks.

- Add a short text note acknowledging Roth’s caveat: pre-trend tests are used descriptively and should not be the sole basis for model selection.

- **Appendix C (Window sensitivity)**

- If you enable it, use it to show alternative windows and possibly **TEU-weighted averages**, referencing BJS and C&S for motivation.

---

7. R^2: what to report and how to interpret it

R^2 is not the star of a DiD/event-study design, but it is standard to report one measure of model fit. In high-dimensional fixed-effects regressions:

- There are often multiple R^2 notions (overall, within, between, adjusted, etc.).
- The **within R^2** — the fraction of **within-unit (terminal) variance** in ln(LP) explained by the regressors — is the most interpretable in your context.

Recommendations:

1. **Report only one R^2 per column:**
2. Use **within R^2** from your fixed-effects estimator (e.g., from `feols` / `reghdfe` / `fixest`).

3. Label it explicitly as "Within R^2" in the row.

4. **Interpretation in text:**

5. R^2 here mainly tells the reader that your model explains a substantial share of variation in ln(LP) after controlling for terminal and month FE.

6. Don't over-interpret differences in R^2 across columns; they are usually modest and driven by adding trends/shocks.

7. **No need for multiple R^2 rows:**

8. Skip overall or between R^2 in the tables; they add clutter without clarifying the treatment effect.

---

8. How to talk about the coefficients in the main text

Putting everything together, here's the **narrative template** you can use when writing your results section, in a way that is consistent with the four papers:

1. **Define the estimand and link to βs**

2. "Our main parameter of interest is the average effect of the port reforms on terminal-level log labor productivity over the first year after the reform. Formally, this is the simple average of event-time coefficients β(m) for m = 1,…,12 in an event-study design with not-yet-treated controls."

3. **Report and interpret the main β**

4. "In Haifa, the entrant terminal (Bayport/SIPG) exhibits an average first-year effect of $\hat{\beta}$ = 0.xx (SE = 0.yy), corresponding to an increase in LP of about Z% relative to the counterfactual. The legacy terminal shows a $\hat{\beta}$ = −0.ww, or a W% decline in LP."

5. **Discuss dynamics qualitatively**

6. "Event-study plots indicate that the gains at the entrant terminal build up gradually, with little movement at m = 0–1 and larger increases by m = 6–12. Legacy terminals show the opposite pattern, with modest pre-trends and a deterioration in LP after the reforms."

7. **Address pre-trends cautiously**

8. "Pre-treatment event-time coefficients are small and imprecise; joint tests of the leads do not reject equality to zero at conventional levels. Following Roth (2024), we interpret this as suggestive (but not definitive) evidence against strong pre-trends."

9. **Summarise robustness**

10. "Alternative specifications that add port-specific trends, control for COVID and war-related shocks, restrict to balanced entrant/legacy samples, or exclude COVID/war windows deliver similar first-year effects, both in sign and magnitude. This is consistent with the robustness principles in recent event-study work (Callaway & Sant'Anna; Borusyak–Jaravel–Spiess)."

11. **Optional: throughput-weighted interpretation**

12. “When we re-weight terminals by throughput (TEU), the estimated first-year gains for entrants become slightly larger, indicating that capacity-weighted productivity improvements are at least as strong as the simple average effects.”

---

9. Summary

- **Callaway & Sant’Anna** give you the conceptual backbone: think in terms of ATT(g, t) and its aggregations; your "Average post m ∈ [1, 12]" is a legitimate dynamic ATT aggregation when defined explicitly.
- **Sun & Abraham** guide the **dynamic presentation**: clearly separated leads and lags, selected horizons plus window averages, and careful interpretation of pre-trend tests.
- **Borusyak, Jaravel & Spiess** motivate your use of **window-average effects**, support diagnostics (N(m)), and window-sensitivity checks.
- **Roth** shapes your **pre-trend narrative** so that you use F-tests and lead coefficients descriptively, not as a hard pass/fail filter.

Applied to your thesis, this means: keep your main tables focused on **average post m ∈ [1, 12]** and **implied %ΔLP**, accompanied by **pre-trend p-values, observations, and within R^2**; use your figures and appendices to show the full dynamic β(m) profiles, pre-trend diagnostics, and robustness across weighting schemes and windows. This will make your empirical presentation look very much in line with contemporary best practice in the DiD/event-study literature.