perplexity

Okay, so for my research, I want to test the causal chain that regime change (i.e., democracy levels represented as DI score), specifically democratic backsliding (period of decrease in DI), disrupts countries' data reporting/quality (i.e. statistical capacity represented as SPI score), and as a result, this disruption to data collection and reporting (decrease in SPI) negatively impacts sustainable development (represented as SDG score). So DI → SPI → SDG. I believe this calls for a mediation / two-staged model structure.

With everything we discussed in this trend, how would you structure my research design? Please devise a comprehensive (but easy to read) outline of my research approach. Keep in mind that I am interested in applying the following quantitative techniques: pooled ols, fixed effects, first difference, mediation and possibly crossed-lagged SEM modeling. Also make sure to include necessary robustness checks wherever is best for organizational purposes.

Excellent. It sounds like you have a clear and powerful causal story to test. Building a research design around a mediation hypothesis (`DI -> SPI -> SDG`) with panel data is a fantastic project.

Based on everything we've discussed, here is a comprehensive research design outline. It's structured to be a logical progression, where each step builds upon the last to create a convincing argument for your final results.

Research Design: A Step-by-Step Approach to Testing the DI → SPI → SDG Causal Chain

Part I: Theory, Hypotheses, and Data Preparation

1. **State the Overarching Causal Theory:** The core of your research is to test the causal chain where democratic backsliding degrades a nation's statistical infrastructure, which in turn hampers its ability to make progress on sustainable development goals.
2. **Formalize the Hypotheses (Two-Stage Mediation):**
    - **Stage 1 Hypothesis:** Democratic backsliding (a decrease in `DI` score) has a negative causal effect on a country's statistical capacity (`SPI` score) in the subsequent period.
    - **Stage 2 Hypothesis:** A lower level of statistical capacity (`SPI` score) has a negative causal effect on a country's SDG progress (`SDG` score) in the subsequent period.

3. **Data Preparation and Descriptive Analysis:**
    - **Variable Operationalization:** Clearly define and source your variables: `DI`, `SPI`, `SDG`, and your primary control variables (`Log(GDP per-capita)`, `GNI_classification`, `Year` dummies).
    - **Descriptive Statistics:** Before any modeling, present a summary table (means, standard deviations, min/max) for all variables. Show a correlation matrix to get a preliminary sense of the relationships.
    - **Data Visualization:** Create time-series plots for your key variables (`DI`, `SPI`, `SDG`) for a few example countries to visually inspect trends.

Part II: The Core Two-Stage Panel Regression Analysis

This is the heart of your study. You will perform this entire sequence for **both stages** of your mediation analysis.

A. Stage 1 Analysis: The Effect of Democracy on Statistical Capacity (`DI` → `SPI`)

1. **Step 1: The "Naive" Baseline - Pooled OLS**
    - **Action:** Run a Pooled OLS regression.
    $$\text{SPI}_{it} = \beta_0 + \beta_1 \text{DI}_{i,t-1} + \beta_2 \text{Controls}_{it} + \epsilon_{it}$$
    - **Purpose:** To establish a simple baseline correlation, ignoring the panel structure. This is the result you will later show is likely biased.
    - **Robustness Check:** Use **country-clustered standard errors**. This is the only check needed at this stage.
2. **Step 2: Accounting for Country Differences - Panel Models**
    - **Action:** Run both a **Random Effects (RE)** and a **Fixed Effects (FE)** model.
    - **Purpose:** To control for the unobserved, time-invariant characteristics of each country (e.g., culture, geography) that likely bias the Pooled OLS results. The FE model is generally considered more robust as it makes fewer assumptions[1] [2].
3. **Step 3: Justifying Your Model Choice - The Hausman Test**
    - **Action:** Perform a Hausman test comparing the RE and FE models.
    - **Purpose:** To formally test whether you can use the more efficient RE model or if you must use the more robust FE model. If the test is significant ($p < 0.05$), you have statistical evidence that unobserved country effects are correlated with your variables, and you must choose the FE model[2].
4. **Step 4: The Final Model and Its Robustness Checks**
    - **Action:** Present your chosen model (likely Fixed Effects) as your main result for Stage 1.
    $$\text{SPI}_{it} = \beta_1 \text{DI}_{i,t-1} + \beta_2 \text{Controls}_{it} + \alpha_i + \lambda_t + \epsilon_{it}$$
    - **Interpretation:** Interpret the size and significance of $ \beta_1 $. This is the estimated effect of democracy on statistical capacity, controlling for all stable country characteristics.
    - **Full Robustness Checks (Apply these to your chosen model):**

- **Serial Correlation:** Use a Wooldridge test to check for autocorrelation in the errors.
- **Heteroskedasticity:** Use a panel-version of the Breusch-Pagan test.
- **Standard Errors:** Regardless of the test results, present your final model with **robust standard errors clustered by country** to handle both issues.
- **Multicollinearity:** Check the Variance Inflation Factor (VIF) for your time-varying controls.

B. Stage 2 Analysis: The Effect of Statistical Capacity on SDG Progress (`SPI` → `SDG`)

- **Action:** Repeat the exact same four steps (Pooled OLS → RE/FE → Hausman Test → Final FE Model with Robustness Checks) for the second stage of your mediation.
- **The Final Model Equation will be:**

$$\text{SDG}_{it} = \beta_1 \text{SPI}_{i,t-1} + \beta_2 \text{DI}_{i,t-1} + \beta_3 \text{Controls}_{it} + \alpha_i + \lambda_t + \epsilon_{it}$$

Note: It is crucial to include the initial variable (`DI`) as a control in this second stage to see if `SPI` mediates its effect.

Part III: Advanced Models for Deeper Causal Inference

These models can be presented as further evidence to strengthen your conclusions.

1. **First-Difference (FD) Model:**
   - **Action:** Re-run your final FE models using the FD specification.
   - **Purpose:** The FD model is an alternative way to eliminate country-specific fixed effects by regressing the *change* in Y on the *change* in X[1]. If the results are very similar to your FE results, it increases confidence in your findings. FE is generally preferred when you have more than two time periods[1].
2. **Cross-Lagged Panel Model using SEM (Structural Equation Modeling):**
   - **Action:** Specify a cross-lagged model that estimates both causal pathways simultaneously[3].
   - **Purpose:** This is your most advanced test. It allows you to:
     - Model the entire `DI → SPI → SDG` chain in a single system.
     - Control for the "autocorrelation" of each variable (e.g., `DI_t1` predicting `DI_t2`).
     - Test for **reciprocal causation** (e.g., does better SDG progress also improve democracy?).
   - **Interpretation:** You will interpret the path coefficients in your model to see if the hypothesized causal links hold up in this more complex and dynamic specification.

Part IV: Synthesizing Results and Drawing Conclusions

1. **Test for Mediation:** Based on your results from the two-stage FE analysis, determine if mediation occurred.
   - Is the effect of `DI` on `SPI` (Stage 1) significant?
   - Is the effect of `SPI` on `SDG` (Stage 2) significant?

    - Does the effect of `DI` on `SDG` become smaller or insignificant when `SPI` is included in the Stage 2 model?
2. **Build Your Argument:** Structure your conclusion by starting with the simple (but biased) OLS results, showing how they change once you apply more rigorous FE/FD models, and finishing with the nuanced story from the cross-lagged SEM. This narrative approach makes for a very powerful and convincing research paper.

⁂

1. TS-Num-4-First-Diff-FE.pdf
2. TS-Num-5-RE-FE-int.pdf
3. TS-Num-6-Survival-Cross-Lag.pdf