## Confidential Comments to Editors

With a clear research question and extensive empirical analysis, the paper demonstrates its potential. However, there is considerable room for improvement, as noted in my comments to the authors. This is the main reason why I recommend rejection of the paper, and I hope the authors address the main concerns and seek the next opportunity.

## Comments to the Authors

This paper builds on a rich literature on the effect of democracy on economic growth by identifying the key institutional building blocks that drive the relationship and exploring the heterogeneity of the effect across countries and time. The expected contribution is to provide an empirical analysis that is far more extensive than existing studies.

The estimation strategy employed in the paper can be summarized as follows:

1. It uses different measures of democracy indicators based on the theory proposed in Section 2 and compares the estimates of the effect of each treatment on economic growth.
2. To estimate the effect, the paper uses a parametric regression model based on a panel structure following Chan and Kwok (2022). This model can be interpreted as an interactive fixed effects model where one may control for unmeasured confounders using the interaction term of unit fixed effects and time fixed effects ($\delta\_i^\prime \hat{f}\_t$). Specifically, the paper first estimates time fixed effects (factors) via PCA using control group data, and then fits the model (Eq. 4) using the entire data set to estimate the regression coefficient of the democracy indicator. The coefficient is defined for each country over time to capture the individual treatment effect on the treated (ITET) averaged over time.
3. Finally, it reports the smoothed line of this regression coefficient against the years that each country spent in democracy, along with the fitted coefficient and its standard error. As mentioned above, the authors report the results for each definition of treatment.

Notwithstanding its clear motivation and objective, the paper has several issues that need to be addressed concerning its identification and estimation strategy.

\*Causal Mechanism of Institutional Building Blocks\*

The most critical limitation of the empirical analysis is that the paper does not clearly explain the causal mechanism of the institutional building blocks. The mechanism proposed in Figure 1 presents the basic components of liberal democracy and their hierarchy but does not fully explain the causal relationships between each component. This leaves the reader unclear how one can interpret the results, especially in the ``horse races’’ between the low-level democracy indicators. These may be prone to post-treatment bias if one of the other components controlled for is a post-treatment variable. The analysis would have been clearer if the authors had presented a Directed Acyclic Graph (DAG), which is widely used in causal inference, and/or elaborated on the causal mechanism (e.g., component A is a moderator/mediator/confounder of B and the outcome).

\*Identification Strategy using Panel Structure\*

The paper carefully explains the identification and estimation strategy, mainly focusing on the endogeneity issue (namely allowing non-parallel trends using interaction terms). However, it is not straightforward why the authors chose the specific method of PCDID among other possible specifications of interactive fixed effects models. In particular, given that the authors are interested in exploring the heterogeneity of the effect in terms of the length of treatment, generalized SCM (Xu 2018), which assumes a similar model and estimation strategy yet estimates the individual treatment effect for each time period (i.e., $\beta\_{it}$), may have been a better fit for this study.

To be more precise, the smoothed line reported in the paper explores the ITET averaged over time (y-axis) across each treated country with different lengths of treatment (x-axis). Thus, it is a bit concerning that each point in the figure corresponds to conditional ATET, where the conditioning variable is the total length of treatment, which is essentially a post-treatment variable (+ mean/sd of control variables if any, as in ``horse races’’). By using generalized SCM, one may explore ITET at each post-treatment period as well as ATET or conditional ATET using pre-treatment variables. It may have been more rigorous, for example, to report the ATET over time to show the heterogeneity in terms of time. Additionally, the authors could specify important pre-treatment variables that may contribute to the heterogeneity of ITET across countries based on theory (i.e., moderators) and implement subgroup analysis to further explore heterogeneity across countries.

\*Minor Comments\*

An important identification assumption of the given model is strict exogeneity, yet it has not been mentioned in the paper. Note that strict exogeneity assumes that the potential outcome is independent of treatment at any point, conditioning on the entire histories of observed covariates and latent fixed effects, and thus does not allow for feedback effects from past outcomes (or covariates) to current and future treatment. Additionally, given that the paper fits multiple models, the authors may consider applying multiple testing corrections. Lastly, a visualization of treatment status (perhaps in the supplementary material) may be beneficial for a visual inspection of outliers and missing values.

This paper starts from an observation that the existing literature on policy evaluation focuses on either a single policy (thereby lacking generalizability) or a set of policies as a whole, i.e., ``policy regime'' (thereby lacking interpretability). Instead, this paper proposes a new approach that sets the unit of analysis as a bundle of policies, or what the paper calls ``policy portfolio''. Specifically, the paper proposes a two-step procedure, where the first step is to construct a high-dimensional data of policy portfolios based on their coding scheme of target and instrument axes, and the second step is to fit a linear model with a regularization based on a strong prior that the effect of the policy portfolio is sparse. Notwithstanding its clear motivation and contribution, the paper has several issues that need to be addressed.

\*Literature Review and Problem Statement\*

The paper will be more convincing if it provides a more comprehensive literature review on policy evaluation. For example, there is a growing literature on policy learning and evaluation using causal inference methods, which are only briefly mentioned in the paper for one strand of the literature that studies a single policy.

On a related note, bringing causal inference methods and framework more clearly into the paper will help readers better understand the main problem that the paper is trying to address. The current problem statement sounds too general; it can be applied to any study not particular to the policy evaluation, which can be improved by making more theoretical discussion. For instance, external validity in causal inference has been a long-standing issue (see Egami and Hartman, forthcoming in APSR), and the discussion on generalizability of policy evaluation can be stated more clearly borrowing this framework. Likewise, the discussion on the interpretability of policy evaluations with aggregate approach can be stated more clearly by borrowing the framework of causal mediation analysis and heterogeneous treatment effect estimation. Given previous works, the paper can make a sharper arguments on the limitation of the existing literature: e.g., the issue with discrepancy between the target population and sample that is analyzed, heterogenous mechanism of policy effects that differs across the time and geographic locations, bundled treatments and so on.

\*Details of the Method\*

In terms of the second step of the proposed method, the paper does not provide enough details on the model specification and estimation. For example, it is not clear why the specific regularization method based on a Student's t-distribution is chosen, given other regularization methods such as LASSO, ridge, and other machine learning methods. Also, despite the fact that this is a Bayesian approach, the paper does not provide any details on the estimation procedure such as choice of priors for other parameters like explanatory variables and MCMC algorithm (or any other methods for MAP estimation such as EM algorithm) that is used for estimation.

There are some strengths of adopting a Bayesian approach that the paper can leverage but is not fully utilized. First of all, the paper can be more explicit about the data generating process of the model. Potentially, the paper can be specific about how the policy portfolios are generated (e.g., how the target and instrument are determined interdependently) and incorporate this information into the model (e.g., using a hierarchical model). Second, instead of suggesting a summary statistic using a naive average of association (point estimate of coefficients), the method can generate a more rigorous and flexible summary statistic using the posterior distribution of the parameters (e.g., using a joint posterior predictive distribution of the coefficients). This is particularly useful given the fact that the paper is highlighting the interdependence of policies, which is not captured by the suggested summary statistic. Lastly, the paper can be more explicit about the uncertainty of the estimation results in the main text.

\*Framing of the Contribution\*

One of the main contributions that is not fully highlighted in the paper is the data collection and coding of policy portfolios in its empirical analysis on ``Porter Hypothesis''. In line with this, and the complexity of the coding scheme of the policy portfolios, it seems to be more promising to frame the paper around this specific empirical application rather than the general methodological contribution. The paper's attempt at providing a guideline on how to code policy portfolios in the first step is helpful. However, given the fact that policy itself is a complex concept that is not easily defined and measured, the concern with measurement error is not fully addressed in the paper. Furthermore, there are remaining issues with measurement, such as whether the two axes of target and instrument are orthogonal (i.e., some target may be more likely to be used with some instrument) so that it does not posit an issue with feasibility, and the fact that a single policy with two different targets and instruments is coded same as two different policies can be problematic given their different level of dependency. On this note, the generalizability of the proposed method is still questionable, and thus emphasizing it hampers the paper's contribution on this specific empirical application with environmental policies.