# Data Suitability Assessment for Quasi-Experimental Building Energy Policy Evaluation

*Hye-Gi Kim, Korea Institute of Civil Engineering and Building Technology*
*Han-Gyeong Chu, Korea Institute of Civil Engineering and Building Technology*
*Ki-Uhn Ahn, Korea Institute of Civil Engineering and Building Technology*
*Sun-Sook Kim, Ajou University*
*Deuk-Woo Kim, Korea Institute of Civil Engineering and Building Technology*

## ABSTRACT

Building energy big data have created valuable opportunities for policy evaluation, yet current use remains limited to descriptive statistics without causal analysis for evidence-based policymaking. Quasi-experimental methodologies, successfully established in medicine, economics, and public administration, offer promising approaches for rigorous policy evaluation. However, applying these methods requires careful assessment of whether available datasets meet specific data requirements.

This research develops a systematic data suitability assessment procedure for applying quasi-experimental methodologies including Difference-in-Differences (DiD), Propensity Score Matching (PSM), Regression Discontinuity Design (RDD), and Interrupted Time Series (ITS) to building energy policy evaluation. The procedure identifies critical data requirements based on data structure characteristics, treatment allocation mechanisms, and covariate availability.

The assessment procedure evaluates datasets across four dimensions: data structure adequacy, methodological applicability, data requirement fulfillment, and implementation feasibility. This structured approach helps researchers determine data suitability for quasi-experimental analysis while identifying gaps requiring additional data collection.

Unlike traditional building-focused Measurement and Verification (M&V) assessments that cannot determine whether policies are effective on average, this approach enables the application of causal inference techniques for evidence-based policy decisions. The procedure facilitates the transition from descriptive energy analysis to causal policy evaluation in the building energy sector. This work provides practical guidance for expanding building energy policy assessment beyond individual buildings to population-level evaluations, helping policymakers and researchers make informed decisions about data collection and analytical approaches based on available data resources.

## Introduction

Rising energy costs and growing concern over energy affordability have moved building energy efficiency to the center of public policy: utility bills strain household and institutional budgets alike, and the public money invested in building energy programs must demonstrably lower those costs to be justified. Climate change has made the building sector a top priority for energy efficiency policies. Buildings account for about 30% of global energy use and 26% of energy-related carbon emissions (IEA, n.d.). To address this challenge, governments worldwide have launched various building energy programs—from energy performance standards and green building certifications to financial incentives and building codes. About 80 countries now have mandatory or voluntary building energy codes in place, representing a 30% increase since the

Paris Agreement in 2015 (IEA, n.d.). Whether these policies work effectively matters greatly, not only for reducing energy use but also for spending public money wisely and meeting national climate goals.

Simultaneously, governments have built large databases that collect building energy information. For example, South Korea's National Building Energy Integrated Management System contains data for over 7 million buildings (Ji et al. 2020). The United States has similar systems, with ENERGY STAR Portfolio Manager tracking data from over 150,000 commercial and multifamily properties (US EPA 2023). These databases have opened valuable opportunities for policy analysis, but realizing that value requires two things. First, the data must be made analytically usable: turning records into causal evidence demands specific analytical approaches and data structures, not mere availability. Second, the data themselves carry selection issues that must be addressed. ENERGY STAR Portfolio Manager, for instance, relies on voluntary benchmarking, so its sample skews toward better-managed buildings; this kind of self-selection must be diagnosed before such data can support credible causal claims.

Despite having these rich datasets, policy evaluators use them in limited ways. Current practice mainly involves providing basic information and creating simple charts: showing average consumption, comparing groups, and displaying trends over time. While these outputs help with monitoring, they cannot answer the key question that policymakers need to know: what is the actual causal effect of the policy? This analytical gap continues even though research from other fields shows that simple comparisons between program participants and non-participants often mislead us. The differences we observe may reflect who chose to participate rather than whether the program actually works (Imbens and Wooldridge 2009).

This gap between available data, what policymakers need to know, and existing methods has created strong demand for Evidence-Based Policy (EBP) approaches. The social sciences—especially economics, public health, and public administration—have made methodological advances over the past twenty years. Researchers have developed and refined quasi-experimental designs for evaluating policies when randomized experiments are not feasible (Angrist and Pischke 2009). DiD, PSM, RDD, and ITS have become standard tools across many policy domains. However, each method imposes specific data requirements—DiD requires panel data with parallel pre-treatment trends, PSM needs rich covariate data, RDD depends on a sharp administrative threshold, and ITS requires long pre/post time series—that may not be obvious to building energy researchers or reflected in existing database designs.

This raises a practical challenge: does my dataset meet the requirements for this type of analysis? Without systematic assessment procedures, researchers may apply methods their data cannot support, or fail to recognize when rigorous causal analysis is already feasible.

This research develops a systematic data suitability assessment procedure for building energy policy evaluation. The procedure evaluates datasets across four dimensions—data structure adequacy, methodological applicability, data requirement fulfillment, and implementation feasibility—providing clear criteria for quasi-experimental method selection and identifying data gaps requiring additional collection.

Unlike traditional M&V protocols, which assess whether energy savings in individual buildings can be measured accurately, this procedure evaluates whether causal policy effects across building populations can be reliably estimated. This distinction matters because population-level effectiveness questions require different data structures and analytical approaches than building-level performance measurement.

The following sections present the assessment procedure, demonstrate it through Korea's Green Remodeling program case study, and conclude with implications for data collection and future research.

## Methods

This section presents the data suitability assessment procedure for quasi-experimental building energy policy evaluation. We begin by defining the scope of our procedure, then review data requirements for four major quasi-experimental methods, and finally describe the assessment procedure organized across four evaluation dimensions.

### Scope and Focus

This research focuses on binary or discrete policy interventions—situations where buildings either receive treatment or do not, or where treatment occurs at discrete time points. While continuous treatment variables (insulation levels, equipment efficiency ratings, incremental technology adoption) are also important policy levers worthy of causal analysis, we limit our current scope to binary/discrete cases to establish foundational assessment criteria. This choice reflects practical considerations: binary treatments allow clearer counterfactual definition and more straightforward application of traditional quasi-experimental methods. Future research should extend the assessment procedure to continuous treatment contexts, which require methodological considerations such as dose-response modeling and generalized propensity scores.

Our procedure examines quasi-experimental methods established in econometrics and program evaluation research: Difference-in-Differences, Propensity Score Matching, Regression Discontinuity Design, and Interrupted Time Series. These methods share a common goal of estimating population-level average treatment effects from observational data when randomized experiments are infeasible. They excel at answering the fundamental policy question: does this intervention work on average across the target population? These four were selected because they are the most widely applied designs in the program-evaluation literature. Broader methodological taxonomies (Shadish, Cook, and Campbell 2002; Handley et al. 2018) catalog additional variants, including instrumental variables and synthetic control; we focus on DiD, PSM, RDD, and ITS because they dominate applied policy evaluation, owing to their transparent identifying assumptions and well-established diagnostic toolkits.

We acknowledge that machine learning approaches—including causal forests, double machine learning, and algorithmic methods—offer valuable capabilities. Machine learning excels at identifying conditional average treatment effects, capturing nonlinear relationships, and making individualized predictions. Traditional quasi-experimental methods, by contrast, focus on estimating aggregate policy effects and providing clear interpretable evidence about whether policies work on average—answering whether a policy achieves its intended aggregate impact, which is the primary concern for policymakers seeking to justify budgets or decide on program continuation. Additionally, traditional methods provide clearer causal interpretation through well-understood identifying assumptions that researchers can assess and communicate to non-technical stakeholders. For these reasons, we focus our data assessment procedure on established quasi-experimental methods; future extensions to machine learning approaches for building energy analysis would be valuable but fall outside our current scope.

## Quasi-Experimental Methods and Data Requirements

Each quasi-experimental method relies on specific data structures and assumptions to identify causal effects. Understanding these requirements helps researchers evaluate whether their available data can support rigorous policy analysis. Table 1 compares the four quasi-experimental methods across key dimensions

Table 1. Comparison of Four Quasi-Experimental Methods

| Method | Data Structure | Key Assumptions | Building Energy Challenges |
|---|---|---|---|
| DiD | Panel (repeated observations) | Parallel trends, no anticipation, SUTVA | Short pre-periods, staggered adoption, weather confounding |
| PSM | Cross-sectional or panel | Unconfoundedness, overlap, positivity | Missing covariates, limited overlap, selection bias |
| RDD | Cross-sectional near threshold | No manipulation, continuity | Fuzzy compliance, small sample near cutoff |
| ITS | Long time series | No contemporaneous shocks, stable trends | Seasonality, autocorrelation, limited time points |

Note: SUTVA = one unit's treatment doesn't affect others. Parallel trends = similar trajectories absent intervention. Unconfoundedness = all confounders observed. Overlap/positivity = treated and untreated units exist across covariate range. No manipulation = units can't precisely control assignment. Continuity = smooth outcome variation around threshold. Fuzzy compliance = imperfect rule adherence. Contemporaneous shocks = simultaneous confounding events.

**Difference-in-Differences (DiD).** DiD compares outcome changes between treated and control buildings over time (Table 1). The method requires panel data with pre/post observations for both groups and assumes parallel trends—that without the policy, both groups would follow similar trajectories (Wing et al. 2018; Angrist and Pischke 2009). For example, evaluating a green building incentive program would compare energy use changes in participating buildings versus non-participating controls. The main challenge is confounding: building outcomes vary with weather, occupancy, and economic conditions, so DiD models must control for these time-varying factors (Wing et al. 2018)

**Propensity Score Matching (PSM).** PSM creates comparable groups by matching buildings with similar characteristics (Table 1). The method estimates propensity scores—the probability of treatment given observed covariates—then matches treated and control units with similar scores (Austin 2011; Stuart 2010). Success depends on measuring covariates affecting both treatment assignment and outcomes—building age, size, occupancy, construction quality, and baseline performance levels. For example, Korea's green remodeling program attracts older, less efficient buildings. Without matching on these characteristics, simple comparisons would conflate program effects with pre-existing differences. The key challenge is unobserved confounders: if factors (like occupant behavior or maintenance quality) aren't measured, matches remain imperfect and bias persists (Ho et al. 2007; Austin 2011).

**Regression Discontinuity Design (RDD).** RDD exploits policy thresholds to estimate effects by comparing units just above versus just below a cutoff (Imbens and Lemieux 2008). The method assumes outcomes vary smoothly around the threshold absent treatment, so discontinuous jumps at the cutoff identify causal effects. Examples include building area cutoffs

(energy codes apply above 1,000 m²) or performance thresholds (buildings above 200 kWh/m²/year receive incentives). The challenge is fuzzy compliance: buildings near thresholds sometimes avoid or manipulate classification (Imbens and Lemieux 2008). Additionally, small sample sizes near cutoffs reduce statistical power, particularly for voluntary programs where few buildings cluster at thresholds.

**Interrupted Time Series (ITS).** ITS detects whether a policy broke existing outcome trends by analyzing time series data before and after intervention (Kontopantelis et al. 2015). The method works well for city-wide building codes or national standards affecting all buildings simultaneously, making control groups unavailable. ITS requires long time series (8-12+ periods pre/post) to establish stable baselines and detect real changes (Kontopantelis et al. 2015). For building energy, main challenges include seasonality (outcomes vary monthly), autocorrelation (this month predicts next month), and limited historical data—many databases lack sufficient pre-policy observations to test trend assumptions.

To develop a systematic assessment procedure, we conducted a comprehensive literature review of data requirements for causal inference in quasi-experimental research. Through systematic review of methodological literature across economics, public health, and program evaluation fields, we identified 22 specific assessment items organized across four dimensions. Table 2 presents these items with their evaluation criteria and supporting references

Table 2. Data Suitability Assessment Items and Evaluation Criteria

| Dim | # | Item | Evaluation Criteria |
|---|---|---|---|
| D1 | 1 | Temporal dimension | Panel data enables before-after comparison and trend analysis (Cham et al. 2024) |
| | 2 | Pre-post availability | Sufficient pre/post periods reduce bias and enable parallel trends testing (Harris et al. 2006; Barnett et al. 2005) |
| | 3 | Treatment timing clarity | Clear intervention timing enables accurate measurement (Schweizer et al. 2016) |
| | 4 | Treatment variation | Variation in timing/intensity enables differential effect analysis (Steiner et al. 2017) |
| | 5 | Control group existence | Controls enable comparison and isolate intervention effects from exogenous factors (Schweizer et al. 2016; Handley et al. 2018) |
| | 6 | Parallel trends testability | Pre-treatment trend similarity verifies counterfactual validity (Wing et al. 2018) |
| | 7 | Covariate sufficiency | Comprehensive covariates control confounding and improve matching (Steiner et al. 2017; Austin 2011) |
| | 8 | Common support adequacy | Propensity score overlap between treated and control groups enables PSM execution; measured by common support ratio (Austin 2011; Stuart 2010) |
| | 9 | Participation mechanism | Treatment assignment mechanism determines degree of self-selection bias; voluntary enrollment implies greater confounding requiring statistical adjustment (Barnett et al. 2005; Shadish et al. 2002) |
| | 10 | Data quality | Reliable measurement strengthens statistical power and interpretation (Wing et al. 2018; Basu et al. 2023) |

| | | | |
|---|---|---|---|
| D2 | 11 | DiD assumptions | Parallel trends, common shocks, and no anticipation effects (Wing et al. 2018; Basu et al. 2023) |
| | 12 | RDD feasibility | Sharp discontinuity with no manipulation around threshold (Steiner et al. 2017; Imbens and Lemieux 2008) |
| | 13 | ITS stability | Sufficient time points and stable trends enable change detection (Schweizer et al. 2016; Kontopantelis et al. 2015) |
| | 14 | PSM quality | Evaluates plausibility of the Conditional Independence Assumption (CIA): adequate observed covariates minimize threat from unobserved confounders (Austin 2011) |
| D3 | 15 | Statistical power | Adequate sample size enables detection of meaningful effects (Dong and Maynard 2013; Batistatou et al. 2014) |
| | 16 | Outcome reliability | Low coefficient of variation ensures stable results (Reed et al. 2002) |
| | 17 | Panel balance | Minimal attrition maintains consistency and statistical efficiency (Cook and Campbell 1979) |
| | 18 | Generalizability | Representative sample supports external validity (Cham et al. 2024; Austin 2011) |
| D4 | 19 | Technical difficulty | Method complexity determines required expertise (Kontopantelis et al. 2015) |
| | 20 | Computational efficiency | Processing requirements affect feasibility for large datasets (Ho et al. 2007) |
| | 21 | Interpretability | Clear results facilitate stakeholder communication (Abadie et al. 2010) |
| | 22 | Robustness testing | Sensitivity analysis capabilities verify result stability (Imbens and Lemieux 2008; Lopez and Gutman 2017) |

Note: D1=Data Structure Adequacy, D2=Methodological Applicability, D3=Data Requirement Fulfillment, D4=Implementation Feasibility. All references correspond to sources in bibliography.

**Dimension 1: Data Structure Adequacy.** This dimension examines fundamental data structure requirements through 10 items. Items 1-2 assess temporal structure: temporal dimension existence evaluates longitudinal tracking capability essential for before-after comparison (Cham et al. 2024), while pre-post availability ensures sufficient observation periods (Harris et al. 2006). Items 3-4 address treatment characteristics: timing clarity enables precise intervention definition (Schweizer et al. 2016), and treatment variation assesses diversity in implementation patterns (Steiner et al. 2017). Items 5-6 evaluate comparison groups: control group existence verifies counterfactual availability (Handley et al. 2018), while parallel trends testability examines assumption verifiability (Wing et al. 2018). Items 7-8 concern confounding control: covariate sufficiency evaluates available control variables (Austin 2011), and matching feasibility assesses balance achievability (Stuart 2010). Items 9-10 address validity: selection bias risk identifies self-selection threats (Shadish et al. 2002), while data quality evaluates measurement reliability (Cook and Campbell 1979).

**Dimension 2: Methodological Applicability.** This dimension assesses method-specific assumption compatibility through 4 items. Item 11 evaluates DiD assumption satisfaction,

particularly parallel trends validity and no anticipation effects (Wing et al. 2018; Basu et al. 2023). Item 12 assesses RDD feasibility, examining discontinuity sharpness and manipulation absence (Imbens and Lemieux 2008). Item 13 evaluates ITS stability, verifying adequate time points and trend stationarity (Kontopantelis et al. 2015; Schweizer et al. 2016). Item 14 examines PSM quality, assessing propensity score overlap and covariate balance achievability (Austin 2011).

**Dimension 3: Data Requirement Fulfillment.** This dimension inventories necessary data characteristics through 4 items. Item 15 evaluates statistical power through sample size adequacy for detecting meaningful effects (Dong and Maynard 2013; Batistatou et al. 2014). Item 16 assesses outcome reliability through measurement precision and coefficient of variation (Reed et al. 2002). Item 17 examines panel balance via attrition rates and missing data patterns (Cook and Campbell 1979). Item 18 evaluates generalizability through sample representativeness and external validity (Cham et al. 2024).

**Dimension 4: Implementation Feasibility.** This dimension addresses practical implementation concerns through 4 items. Item 19 evaluates technical difficulty through required expertise and methodological complexity (Kontopantelis et al. 2015). Item 20 assesses computational efficiency via processing requirements and scalability (Ho et al. 2007). Item 21 examines interpretability through result clarity and stakeholder accessibility (Abadie et al. 2010). Item 22 evaluates robustness testing feasibility via sensitivity analysis and specification testing capabilities (Imbens and Lemieux 2008; Lopez and Gutman 2017).

## Data Suitability Assessment Procedure

The assessment procedure evaluates datasets across four dimensions with 22 specific items (Table 3). Each item uses a 5-point scale with quantitative criteria tailored to building energy policy contexts. The four dimensions are weighted based on their relative importance for valid causal inference:

D1 receives the highest weight (35%) as it establishes foundational prerequisites: without appropriate data structure, no quasi-experimental method can proceed. D2 receives 25% as method-specific assumptions determine validity but can sometimes be addressed through design adjustments. D3 and D4 each receive 20% as they represent important but potentially addressable constraints.

Because few evaluators will have all 22 data elements on hand, the assessment is best read as two sequential questions. The first is whether any causal analysis is feasible at all: this requires a clearly defined intervention with known timing (Items 3, 4), a measured outcome aligned with that intervention, and at least one source of identifying variation, namely an untreated comparison group, a sufficient pre/post time series, or a sharp assignment threshold. If none of these identifying strategies is available, no quasi-experimental method can recover a causal estimate and the assessment stops at this stage. The second question, asked only once feasibility is established, is which method is most suitable: each candidate method has a gate item that rules it out when unmet (final column of Table 5). The gate items act as non-compensatory screens: a method stays eligible only when the design feature it depends on is actually present in the data, and a method whose gate feature is missing is excluded regardless of its other scores. The weighted dimension scores are compensatory and serve only to rank the methods that remain, so the item weights express relative importance for ranking, not feasibility. This two-stage logic is why the absence of a sharp assignment threshold (Item 12) eliminates

RDD, whereas the lack of a comparison group (Item 5) eliminates DiD and PSM but not interrupted time series. Figure 1 summarizes this two-stage decision logic.

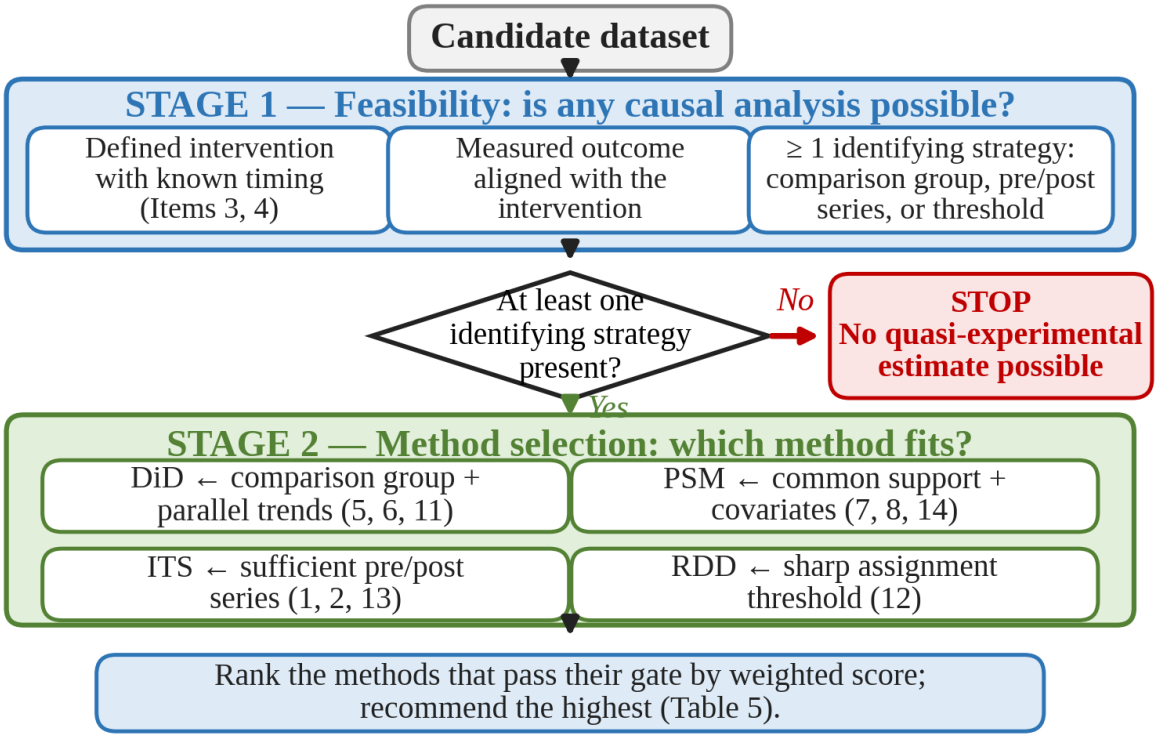


*Figure 1. Two-stage decision logic of the assessment: a feasibility screen (Stage 1) followed by method selection among DiD, PSM, ITS, and RDD (Stage 2).*

Table 3. Assessment Checklist with Scoring Examples

| Dim | Item | Wt% | 5-Point Scale Example |
|---|---|---|---|
| D1 | 1. Temporal dimension existence | 4 | 5=≥48 months; 4=24-47mo; 3=12-23mo; 2=6-11mo; 1=<6mo |
| | 2. Pre-post data availability | 4 | 5=≥12mo each; 4=≥6mo; 3=≥3mo; 2=≥1mo; 1=none |
| | 3. Treatment timing clarity | 3 | 5=precise date; 4=month; 3=quarter; 2=year; 1=unclear |
| | 4. Treatment variation | 3 | 5=high variation in timing and intensity; 4=moderate-high; 3=moderate; 2=low variation; 1=none |
| | 5. Control group existence | 4 | 5=large well-matched control pool; 4=good controls; 3=acceptable; 2=limited controls; 1=none |
| | 6. Parallel trends testability | 4 | 5=≥4 pre-periods; 4=3 pre-periods; 3=2 pre-periods; 2=1 pre-period; 1=untestable |
| | 7. Covariate sufficiency | 4 | 5=comprehensive (≥10 covariates); 4=good (7–9); 3=adequate (4–6); 2=limited (2–3); 1=minimal (≤1) |
| | 8. Common support adequacy | 4 | 5=PS overlap≥90%; 4=75–90%; 3=60–75%; 2=40–60%; 1=<40% |

| | | | |
|---|---|---|---|
| | 9. Participation mechanism | 3 | 5=mandatory assignment; 4=strong administrative incentive; 3=financial incentive/voluntary; 2=weak incentive; 1=fully voluntary |
| | 10. Data quality | 2 | 5=verified, <1% missing; 4=good, 1–5% missing; 3=standard, 5–15%; 2=poor, 15–30%; 1=questionable, >30% |
| D2 | 11. DiD assumption satisfaction | 6 | 5=parallel trends verified, no anticipation; 4=largely plausible, minor concerns; 3=plausible but unverified; 2=trends questionable; 1=violated |
| | 12. RDD feasibility | 6 | 5=sharp cutoff, no manipulation; 4=sharp, minor compliance issues; 3=fuzzy but usable; 2=fuzzy, significant issues; 1=no threshold |
| | 13. ITS stability | 6 | 5=≥12 periods each; 4=10–11; 3=8–9; 2=5–7; 1=<5 |
| | 14. PSM quality | 7 | 5=CIA plausible, rich covariates; 4=largely plausible, minor unobserved confounding; 3=moderate confounding risk; 2=high unobserved confounding risk; 1=CIA implausible |
| D3 | 15. Statistical power | 5 | 5=≥500 units; 4=200-499; 3=100-199; 2=50-99; 1=<50 |
| | 16. Outcome reliability | 5 | 5=CV<10%; 4=10-20%; 3=20-30%; 2=30-50%; 1=>50% |
| | 17. Panel balance | 5 | 5=<5% attrition; 4=5-10%; 3=10-20%; 2=20-30%; 1=>30% |
| | 18. Generalizability | 5 | 5=nationally representative; 4=mostly representative; 3=acceptable; 2=limited representativeness; 1=highly biased |
| D4 | 19. Technical difficulty (inverse) | 5 | 5=minimal expertise required; 4=moderate-low; 3=moderate; 2=moderate-high; 1=advanced expertise required |
| | 20. Computational efficiency | 5 | 5=fast processing; 4=moderate-fast; 3=manageable; 2=slow; 1=computationally intensive |
| | 21. Interpretability | 5 | 5=highly interpretable; 4=mostly interpretable; 3=moderate; 2=difficult for non-experts; 1=complex |
| | 22. Robustness testing | 5 | 5=extensive tests possible; 4=multiple tests feasible; 3=basic; 2=limited options; 1=minimal |

Note: D1=Data Structure Adequacy (35%), D2=Methodological Applicability (25%), D3=Data Requirement Fulfillment (20%), D4=Implementation Feasibility (20%). Wt%=individual item weight. CV=coefficient of variation. Scoring criteria derived from systematic literature review and expert validation.

## Results and Discussion

We demonstrate the data suitability assessment procedure through a case study of Korea's Green Remodeling (GR) program applied to childcare centers, then discuss broader implications for building energy policy evaluation.

**Case Study: Korea's Green Remodeling Program for Childcare Centers.** Korea's Green Remodeling (GR) program for public buildings, administered by the Ministry of Land,

Infrastructure and Transport, provides government subsidies for energy-efficiency retrofits in existing public facilities. The program targets buildings more than 10 years old with poor energy performance, prioritizing facilities serving vulnerable populations—including childcare centers, community health posts, senior centers, and libraries. Eligible institutions (public agencies or local governments that own or manage qualifying buildings) submit applications to the GR Center; projects are then evaluated through on-site inspection and scored on urgency and expected energy savings, with final selections made by a review committee. Construction start and completion dates are recorded in the GR database, providing a well-defined intervention timeline suitable for causal evaluation.

***Data Assembly.*** The dataset integrates four administrative sources: the GR program database (Ministry of Land, Infrastructure and Transport), the National Building Energy Integrated Management System (energy consumption, building registration), the publicly accessible childcare center database (Ministry of Education; building type, operational status, capacity), and regional weather data from the Korea Meteorological Administration (monthly CDD and HDD). Figure 2 illustrates the sample construction process.

The treatment group consists of childcare buildings designated as GR program recipients in the 2020 fiscal year. The GR database identified 251 eligible buildings classified as single building in a lot, of which 239 were matched across databases. Sequential filtering removed buildings closed or dormant for more than one year, those lacking complete operational records across 2018–2023, buildings housing more than one childcare center (to ensure 1:1 building-building correspondence), and buildings that changed operational type mid-period (e.g., private to public). After requiring complete monthly electricity consumption records (non-zero values across all 72 months, used as a reporting-completeness screen, since electricity is metered in every facility whereas gas and district heat may legitimately read zero in some months) and valid floor area data, 152 treatment buildings remained for the analysis dataset, whose outcome variable is total energy use, the combined sum of electricity, gas, and district heat.

The control group was constructed from 12,591 childcare buildings initially identified in the National Building Energy Integrated Management System with matching building type. Applying equivalent quality filters—removing closed buildings, requiring complete 2018–2023 operational records, excluding potential GR recipients in 2021–2024 to prevent treatment contamination, removing type-changed buildings, and requiring complete electricity and area records—yielded 4,441 control buildings. Since all 152 treatment buildings are public childcare centers, the control group was accordingly restricted to public childcare centers, yielding 603 control buildings. The final dataset for the full suitability assessment thus encompassed 152 treatment and 603 control buildings.

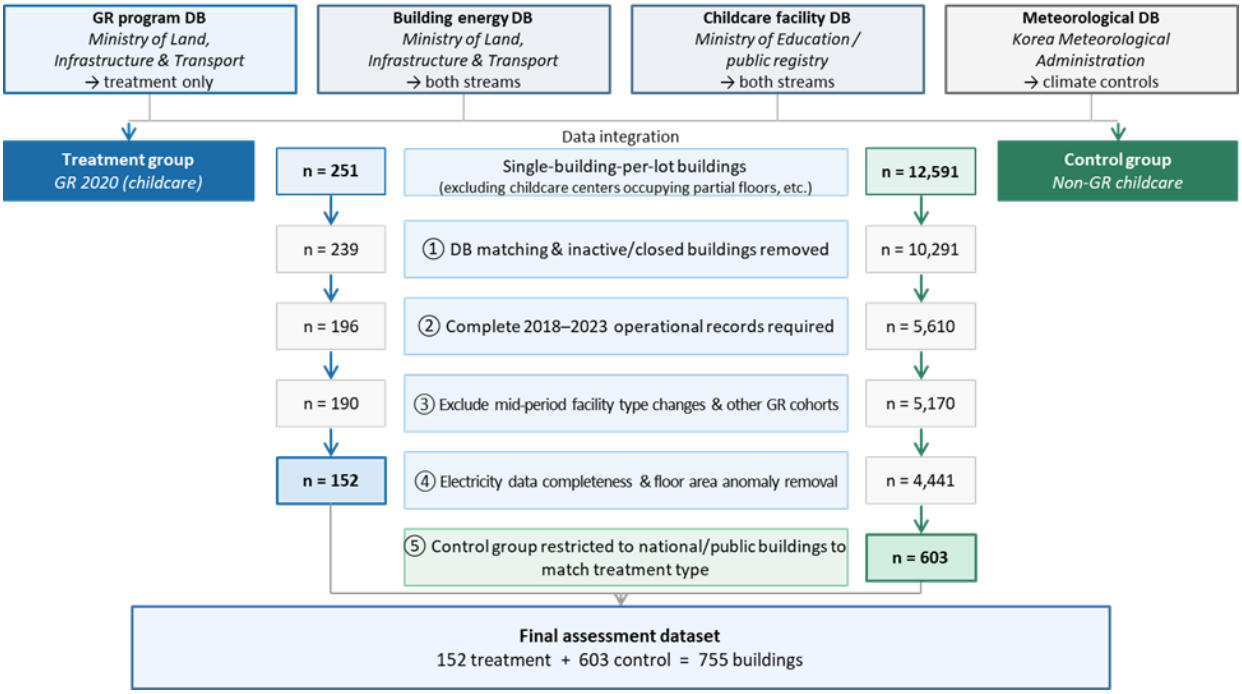


*Figure 2. Data selection process for the childcare center case study. Treatment group (GR 2020 program): 251 initial → 152 final. Control group: 12,591 initial → 4,441 final. Full suitability assessment dataset: 152 treatment + 603 control facilities.*

**Assessment Results.** Table 4 presents scoring results across all 22 assessment items.

Table 4. Data suitability assessment scores for the childcare center case study (treatment: 152 GR-program facilities; control: 603 public childcare centers).

| Dimension | Category | Item# | Assessment Item | Weight | Score | Wt. Score |
|---|---|---|---|---|---|---|
| D1: Data Structure Adequacy (35%) | Temporal structure | 1 | Temporal dimension existence | 0.04 | 5 | 0.20 |
| | | 2 | Pre-post data availability | 0.04 | 5 | 0.20 |
| | Treatment characteristics | 3 | Treatment timing clarity | 0.03 | 5 | 0.15 |
| | | 4 | Treatment variation | 0.03 | 4 | 0.12 |
| | Comparison group quality | 5 | Control group existence | 0.04 | 3 | 0.12 |
| | | 6 | Parallel trends testability | 0.04 | 5 | 0.20 |
| | Confounder control | 7 | Covariate sufficiency | 0.04 | 5 | 0.20 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 8 | Common support adequacy | 0.04 | 4 | 0.16 |
| | Validity threats | 9 | Participation mechanism | 0.03 | 3 | 0.09 |
| | | 10 | Data quality | 0.02 | 3 | 0.06 |
| D2: Methodological Applicability (25%) | DiD applicability | 11 | DiD assumption satisfaction | 0.06 | 3 | 0.18 |
| | RDD applicability | 12 | RDD feasibility | 0.06 | 3 | 0.18 |
| | ITS applicability | 13 | ITS stability | 0.06 | 5 | 0.30 |
| | PSM applicability | 14 | PSM quality | 0.07 | 4 | 0.28 |
| D3: Data Requirement Fulfillment (20%) | Sample size | 15 | Statistical power | 0.05 | 5 | 0.25 |
| | Measurement precision | 16 | Outcome reliability | 0.05 | 1 | 0.05 |
| | Data completeness | 17 | Panel balance | 0.05 | 3 | 0.15 |
| | External validity | 18 | Generalizability | 0.05 | 3 | 0.15 |
| D4: Implementation Feasibility (20%) | Analysis complexity | 19 | Technical difficulty (inverse) | 0.05 | 2 | 0.10 |
| | Computational resources | 20 | Computational efficiency | 0.05 | 2 | 0.10 |
| | Result interpretation | 21 | Interpretability | 0.05 | 2 | 0.10 |
| | Robustness | 22 | Robustness testing feasibility | 0.05 | 5 | 0.25 |

***Dimension 1: Data Structure Adequacy (weighted average 4.29/5).*** Most items scored 4–5, confirming that longitudinal coverage, treatment timing, parallel trends testability, and covariate availability are sufficient for analysis after standard preprocessing. Two items warrant attention. Participation mechanism (Item 9, score 3) reflects voluntary enrollment driven by financial incentives, indicating self-selection: facilities that apply for the GR program likely differ systematically from non-participants in unobserved motivation, management quality, or institutional capacity. This self-selection pattern signals that naive before–after or treatment–control comparisons would conflate program effects with pre-existing facility differences. Common support adequacy (Item 8, score 4) indicates that propensity score overlap between treatment and control facilities falls in the 75–90% range, confirming that matched counterparts are available for most treated facilities.

***Dimension 2: Methodological Applicability (weighted average 3.76/5).*** This dimension assesses how well the data satisfies the identifying assumptions specific to each method. ITS stability (Item 13) scored highest (5): the 72-month window exceeds the minimum threshold for reliable trend estimation. DiD assumption satisfaction (Item 11, score 3) reflects that parallel pre-

treatment trends between GR participants and non-participants are plausible given the administrative nature of the data but require formal testing; anticipation effects prior to program announcement cannot be ruled out. PSM quality (Item 14, score 4) indicates that the Conditional Independence Assumption is largely plausible: ten administrative covariates capture the primary determinants of GR program selection, though unobserved factors such as facility manager motivation remain a residual concern. RDD feasibility (Item 12, score 3) reflects the absence of a single sharp administrative cutoff for program entry; eligibility is based on building age and energy performance criteria applied through a discretionary subsidy allocation process, making a clean regression discontinuity design difficult to justify.

***Dimension 3: Data Requirement Fulfillment (weighted average 3.2/5).*** This dimension inventories whether the dataset meets the quantitative requirements for reliable causal estimation. Statistical power (Item 15, score 5) is fully satisfied: the combined sample of 755 facilities (152 treatment, 603 control) exceeds the 500-unit threshold for moderate effect sizes. Panel balance (Item 17, score 3) reflects 10–20% attrition across the observation window, from facility closures and reporting gaps, which is manageable with appropriate missing data treatment. Generalizability (Item 18, score 3) is moderate: the sample covers multiple administrative regions but overrepresents urban facilities relative to the national childcare center population. The primary constraint is outcome reliability (Item 16, score 1): monthly total energy consumption exhibits a coefficient of variation exceeding 50%, driven by seasonal heating and cooling demand and cross-facility heterogeneity in building size and occupancy. This high variability does not preclude analysis but increases minimum detectable effect sizes and underscores the importance of weather normalization and building-characteristic controls before estimating treatment effects.

***Dimension 4: Implementation Feasibility (weighted average 2.75/5).*** This dimension reflects practical constraints on method implementation. Technical difficulty (Item 19, score 2), computational efficiency (Item 20, score 2), and interpretability (Item 21, score 2) all scored low, reflecting the inherent complexity of quasi-experimental analysis: implementing DiD with staggered treatment timing, running PSM with multiple covariates, and communicating conditional average treatment effects to non-specialist policymakers each require methodological expertise. These scores do not disqualify any method but indicate that institutional capacity must be in place before deployment. Robustness testing feasibility (Item 22, score 5) is the exception: the 72-month time series enables placebo tests, event-study specifications, alternative control group definitions, and bandwidth checks, providing strong post-estimation credibility for all candidate methods.

## Method Selection

Table 5 presents method-specific suitability rankings derived from normalized weighted-average scores on assessment items most relevant to each method. DiD ranked first (normalized weighted score 4.20 across Items 1, 2, 4, 5, 6, 7, 10, 11, 15, 17, and 22), supported by the dataset's strong longitudinal structure, testable parallel trends, and large sample size. PSM ranked second (normalized weighted score 4.05 across Items 5, 7, 8, 9, 14, 15, 18, and 22). Although PSM does not lead the ranking, the voluntary participation design of the GR program means that facilities self-select into treatment based on motivation and institutional capacity, generating systematic pre-treatment differences that DiD alone cannot remove. The program's financial incentive structure (Item 9, score 3), sufficient observable covariates (Item 7, score 5),

and adequate propensity score overlap (Item 8, score 4) confirm that PSM is both necessary and executable. ITS ranked third (normalized weighted score 3.95), benefiting from long time series but penalized by the absence of a dedicated control group and outcome variability. RDD was eliminated: the sharp unmanipulable assignment threshold required for valid RDD application cannot be satisfied by this voluntary program design.

Table 5. Method-specific suitability rankings for quasi-experimental analysis of the Green Remodeling program.

| Method | Relevant Items | Average Score | Recommended Priority | Key requirement (gate item) |
|---|---|---|---|---|
| DiD | 1, 2, 4, 5, 6, 7, 10, 11, 15, 17, 22 | 4.20 | 1 (Recommended) | Parallel trends and a comparison group (Items 5, 6, 11) |
| ITS | 1, 2, 5, 10, 13, 15, 16, 17, 22 | 3.95 | 3 | Sufficient pre/post time series (Items 1, 2, 13) |
| PSM | 5, 7, 8, 9, 14, 15, 18, 22 | 4.05 | 2 | Common support and rich covariates (Items 7, 8, 14) |
| RDD | 3, 7, 12, 18 | — | Not applicable (Item 12 unmet) | Sharp assignment threshold (Item 12); unmet here |

## Conclusions

This paper presents a systematic data suitability assessment procedure for applying quasi-experimental methods to building energy policy evaluation. The procedure organizes 22 assessment items across four dimensions and applies a two-stage logic: a feasibility screen for whether any causal analysis is possible, followed by method selection in which each candidate method must clear a gate—its core identifying requirement—before its weighted score ranks the eligible methods, yielding transparent and reproducible selection. By making this diagnostic step explicit and systematic, the procedure reduces the risk of misapplied methods and supports credible evidence-based policy decisions.

Applied to Korea's Green Remodeling program for childcare centers, the assessment flagged self-selection from voluntary participation (Item 9) and high outcome variability (Item 16), and recommended a combined PSM+DiD design suited to the dataset's structure.

Several limitations point to future work. The scoring criteria reflect building energy policy contexts and may require recalibration for other sectors, and the case study covered a single program year and building type, so replication across cohorts and building categories is needed to establish generalizability. The procedure is building-type agnostic and targets population-level effects rather than individual-building performance. Finally, it identifies data suitability but does not execute the analysis itself; future work will report the PSM+DiD estimates for the Green Remodeling program as an empirical test of the recommendation.

## Acknowledgement

Research for this paper was conducted under the KICT Research Program (project no. 20260130-001, Beyond XAI: Research on Applying Causal Inference for Evaluating the Effects of Building Energy Policies and Technologies) funded by the Ministry of Science and ICT.

## Citations, References, and AI-assisted Tools

### References


Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association* 105 (490): 493–505.

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. “Mostly Harmless Econometrics: An Empiricist’s Companion.” In *Mostly Harmless Econometrics: An Empiricist’s Companion*.

Austin, Peter C. 2011. “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.” *Multivariate Behavioral Research* 46 (3): 399–424. https://doi.org/10.1080/00273171.2011.568786.

Barnett, Adrian G., Jolieke C. van der Pols, and Annette J. Dobson. 2005. “Regression to the Mean: What It Is and How to Deal with It.” *International Journal of Epidemiology* 34 (1): 215–20.

Basu, Sanjay, Ankur Meghani, and Arjumand Siddiqi. 2023. “A Comparison of Quasi-Experimental Methods with Data before and after an Intervention: An Introduction for Epidemiologists and a Simulation Study.” *International Journal of Epidemiology* 52 (5): 1522–37.

Batistatou, Evridiki, Chris Roberts, and Steve Roberts. 2014. “Sample Size and Power Calculations for Trials and Quasi-Experimental Studies with Clustering.” *The Stata Journal* 14 (1): 159–75.

Cham, Heining, Hyunjung Lee, and Igor Migunov. 2024. “Quasi-Experimental Designs for Causal Inference: An Overview.” *Asia Pacific Education Review* 25 (3): 611–27. https://doi.org/10.1007/s12564-024-09981-2.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues in Field Settings*.

Dong, Nianbo, and Rebecca Maynard. 2013. “PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies.” *Journal of Research on Educational Effectiveness* 6 (1): 24–67.

Handley, Margaret A., Courtney R. Lyles, Charles McCulloch, and Adithya Cattamanchi. 2018. “Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research.” *Annual Review of Public Health* 39: 5–25.

Harris, Anthony D., Jessina C. McGregor, Eli N. Perencevich, et al. 2006. “The Use and Interpretation of Quasi-Experimental Studies in Medical Informatics.” *Journal of the American Medical Informatics Association: JAMIA* 13 (1): 16–23. https://doi.org/10.1197/jamia.M1749.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis* 15 (3): 199–236. https://doi.org/10.1093/pan/mpl013.

IEA. n.d. “Buildings - Energy System.” Accessed March 9, 2026. https://www.iea.org/energy-system/buildings.

Imbens, Guido W., and Thomas Lemieux. 2008. “Regression Discontinuity Designs: A Guide to Practice.” *Journal of Econometrics*, The regression discontinuity design: Theory and applications, vol. 142 (2): 615–35. https://doi.org/10.1016/j.jeconom.2007.05.001.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation.” *Journal of Economic Literature* 47 (1): 5–86. https://doi.org/10.1257/jel.47.1.5.

Ji, Chang-Yoon, Min-Seok Choi, Oh-In Gwon, Ha-Rim Jung, and Sung-Eun Shin. 2020. “Greenhouse Gas Emissions from Building Sector based on National Building Energy Database.” *Journal of the Architectural Institute of Korea Structure & Construction* 36 (4): 143–52. https://doi.org/10.5659/JAIK_SC.2020.36.4.143.

Kontopantelis, Evangelos, Tim Doran, David A. Springate, Iain Buchan, and David Reeves. 2015. “Regression Based Quasi-Experimental Approach When Randomisation Is Not an Option: Interrupted Time Series Analysis.” Research Methods & Reporting. *BMJ* 350 (June): h2750. https://doi.org/10.1136/bmj.h2750.

Lopez, Michael J., and Roee Gutman. 2017. “Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas.” *Statistical Science* 32 (3): 432–54.

Reed, George F., Freyja Lynn, and Bruce D. Meade. 2002. “Use of Coefficient of Variation in Assessing Variability of Quantitative Assays.” *Clinical and Vaccine Immunology* 9 (6): 1235–39.

Schweizer, Marin L., Barbara I. Braun, and Aaron M. Milstone. 2016. “Research Methods in Healthcare Epidemiology and Antimicrobial Stewardship–Quasi-Experimental Designs.” *Infection Control & Hospital Epidemiology* 37 (10): 1135–40.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Steiner, Peter M., Yongnam Kim, Courtney E. Hall, and Dan Su. 2017. “Graphical Models for Quasi-Experimental Designs.” *Educational and Psychological Measurement* 77 (5): 916–39.

Stuart, Elizabeth A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science* 25 (1): 1–21.

US EPA, OA. 2023. “EPA Launches Online Tool Providing Energy Use Data and Insights from ENERGY STAR® Portfolio Manager®.” News Release. October 25. https://www.epa.gov/newsreleases/epa-launches-online-tool-providing-energy-use-data-and-insights-energy-starr-portfolio.

Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. “Designing Difference in Difference Studies: Best Practices for Public Health Policy Research.” *Annual Review of Public Health* 39 (Volume 39, 2018): 453–69. https://doi.org/10.1146/annurev-publhealth-040617-013507.

**Disclosure template**

AI-assisted tools used: Claude (Anthropic)
Used for: Copyediting and light editing of manuscript text
Extent: light editing
Human review: The authors reviewed and verified all content, calculations, and citations.