The Challenge of Evaluating Universal School Meals: A Cautionary Tale on Recall-Window Mismatch and Limited Pre-Treatment Data

APEP Autonomous Research*
@"ai1scl"

January 29, 2026

Abstract

Between 2022 and 2023, nine U.S. states adopted universal free school meals policies, providing breakfast and lunch to all public school students regardless of family income. This paper investigates whether these policies reduced household food insecurity beyond direct child beneficiaries through a "resource reallocation" mechanism. Using the Current Population Survey Food Security Supplement (2022–2024), I estimate difference-in-differences models comparing households with school-age children in treatment versus control states. The naive TWFE point estimate of 4.7 percentage points (SE = 2.0 pp, 95% CI: [0.9 pp, 8.5 pp]) on a restricted sample of 2023 adopters versus never-treated states is statistically significant but **meaningless as a causal estimate**—it does not identify a treatment effect because the 12-month recall window does not align with survey-year treatment coding. More informatively, a triple-difference specification with state×year fixed effects comparing households with versus without school-age children yields a precisely estimated null effect (−0.8 pp, SE = 1.3 pp, 95% CI: [−3.4 pp, 1.8 pp]). With only 2–3 years of data, no true pre-treatment periods for 2022 adopters, and policy adoption coinciding with major post-pandemic economic shifts, credible causal inference is not possible with this data structure. This paper contributes a concrete illustration of how recall-window mismatch invalidates standard DiD designs, with implications for researchers using survey data with rolling reference periods.

JEL Codes: I38, H75, C21, C23

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch

Keywords: school meals, food insecurity, difference-in-differences, parallel trends, recall window

1. Introduction

Food insecurity remains a persistent challenge in the United States, affecting approximately 13 percent of all households and 17 percent of households with children in 2023 (Coleman-Jensen et al., 2024). School nutrition programs have long been a cornerstone of the social safety net, providing meals to children from low-income families through the National School Lunch Program (NSLP) and School Breakfast Program (SBP). Beginning in 2022, a wave of state-level reforms went further: California, Maine, Massachusetts, Nevada, and Vermont adopted universal free school meals, extending free breakfast and lunch to *all* public school students regardless of family income. Colorado, Michigan, Minnesota, and New Mexico followed in 2023.

These universal programs represent a substantial expansion of food assistance policy. Unlike traditional means-tested programs, universal meals eliminate the stigma associated with receiving free lunch, simplify school administration, and potentially increase meal participation rates (Gordon et al., 2007; Schwartz and Rothbart, 2020). An open empirical question is whether these programs reduce food insecurity not just for children, but for the entire household through a "resource reallocation" mechanism: when families no longer pay for school meals—even partial copays under reduced-price NSLP—freed resources may improve food provisioning for all household members. Hoynes, Schanzenbach & Almond (2016) document substantial long-run effects of early-life access to food stamps, suggesting that food assistance programs can have broad impacts on household well-being beyond immediate nutritional effects.

This paper attempts to investigate the household-level effects of universal school meals using a difference-in-differences (DiD) design with the Current Population Survey Food Security Supplement (CPS-FSS). However, as I demonstrate, credible causal inference is not possible with the available data structure. The fundamental problem is a mismatch between the treatment indicator—coded at the survey-year level—and the outcome measure—which captures food security over the *prior 12 months* rather than a calendar year aligned with policy timing.

To illustrate what goes wrong, I estimate a standard two-way fixed effects (TWFE) specification despite its known limitations (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021). The coefficient of 4.7 percentage points (SE = 2.0 pp, 95% CI: [0.9 pp, 8.5 pp]) is statistically significant but substantively implausible—it would imply that providing free food to children *increases* household food insecurity. This counterintuitive finding is a red flag of specification failure, not evidence of perverse policy effects. The positive sign reflects confounded timing rather than causal

relationships.

The central contribution of this paper is not about universal school meals per se—it is about the severe limitations of attempting causal inference when (1) the outcome's recall window does not align with treatment timing, (2) pre-treatment data is unavailable or contaminated, and (3) coincident shocks differentially affect treatment and control groups. These challenges arise commonly in applied work using survey data with rolling reference periods, including health surveys with recall-based outcomes, crime victimization surveys, and consumption modules in household surveys.

Three specific identification problems undermine this analysis. First, the CPS-FSS data extract I analyze covers only 2022–2024 (though CPS-FSS is available back to 1995 through IPUMS). This means my sample contains zero pre-treatment observations for the five states that adopted universal meals in August 2022. Without pre-treatment data, it is impossible to assess whether treatment and control states were on parallel trajectories—the core identifying assumption of DiD designs (Roth, 2022; Rambachan and Roth, 2023). Second, the CPS-FSS measures food security over the prior 12 months (approximately December of the prior year through November of the current year for December surveys), while treatment is coded at the survey-year level. For August 2023 adopters, the December 2023 survey captures mostly pre-treatment months (December 2022 through July 2023), with only approximately 4 months of policy exposure (August through November 2023). The treatment indicator thus conflates treated and untreated exposure windows. Third, the study period 2022–2024 coincides with major economic and policy changes including the end of pandemic-era food assistance (SNAP Emergency Allotments ended March 2023, P-EBT phased out), 40-year-high inflation in food prices, and differential state economic recovery from the COVID-19 pandemic.

To stress-test the naive estimates, I implement several robustness checks. A triple-difference specification with state×year fixed effects, comparing households with school-age children to households without children, yields a point estimate of -0.8 percentage points (SE = 1.3 pp, 95% CI: [-3.4 pp, 1.8 pp])—essentially zero and precisely estimated. This specification differences out state-year shocks that affect all households regardless of whether they have school-age children, providing a cleaner comparison. Randomization inference, which permutes treatment assignment across states, yields a p-value of 0.015 for the naive DiD, suggesting the positive coefficient is unlikely under random assignment but not necessarily causal. The discrepancy between the naive DiD (positive, significant) and triple-difference (zero, insignificant) underscores that state-level confounds, not school meal policies, drive the naive results.

The remainder of this paper proceeds as follows. Section 2 develops a theoretical framework for understanding recall-window mismatch in DiD designs, including Monte Carlo

evidence. Section 3 describes the institutional background on universal school meals and the resource reallocation hypothesis. Section 4 presents the data and empirical strategy. Section 5 reports results from TWFE, Callaway-Sant'Anna, and triple-difference specifications. Section 6 discusses why these estimates should not be interpreted causally and what credible identification would require. Section 7 concludes with implications for applied research using recall-based survey outcomes.

2. Theoretical Framework: Recall-Window Mismatch in DiD

Before turning to the institutional setting and empirical analysis, I develop a general theoretical framework for understanding how recall-window mismatch invalidates standard DiD estimands. This framework applies to any setting where (1) the outcome is measured over a rolling reference period and (2) treatment is coded discretely at the survey level.

2.1 Setup and Notation

Consider a panel of states $s \in \{1, \ldots, S\}$ observed at discrete survey times $t \in \{1, \ldots, T\}$. Each survey t measures an outcome Y_{st} that aggregates conditions over a recall window of length L months preceding the survey. For concreteness, assume surveys occur in December and the recall window is 12 months, so survey t captures conditions from December of year $t-1$ through November of year t.

Let τ index months within the recall window. Define the monthly potential outcomes $Y^*_{s\tau}(d)$ for household in state s during month τ under treatment status $d \in \{0, 1\}$. The observed annual outcome is an aggregator of these monthly outcomes:

$$Y_{st} = g\left(\{Y^*_{s\tau}(D_{s\tau})\}_{\tau \in W_t}\right) \tag{1}$$

where W_t denotes the set of months in the recall window for survey t, $D_{s\tau}$ is the actual treatment status in state s during month τ, and $g(\cdot)$ is the aggregation function (e.g., the indicator for "any month food insecure" or the mean monthly food insecurity rate).

State s adopts treatment at calendar time τ_s (the adoption month). For never-adopters, $\tau_s = \infty$. The standard DiD treatment indicator codes state s as treated in survey t if τ_s falls before survey t:

$$\text{Treated}_{st} = \mathbf{1}[\tau_s \leq t_{\text{start}}] \tag{2}$$

where t_{start} is some reference point for survey t (often the calendar year). This binary coding ignores the fraction of the recall window actually exposed to treatment.

2.2 The Exposure Intensity Problem

Define the true exposure intensity for state s in survey t as:

$$E_{st} = \frac{1}{L} \sum_{\tau \in W_t} D_{s\tau} = \frac{\text{Number of treated months in recall window}}{L} \tag{3}$$

For a state adopting in August of year g:

- Survey $t < g$: $E_{st} = 0$ (no treated months)
- Survey $t = g$: $E_{st} = 4/12 \approx 0.33$ (August through November treated)
- Survey $t > g$: $E_{st} = 12/12 = 1$ (full recall window treated)

The binary indicator Treated_{st} codes both survey g and $g + 1$ as "treated = 1," despite exposure intensities of 0.33 and 1.0 respectively. This creates two distinct problems.

Problem 1: Attenuation bias. Under standard parallel trends, the binary TWFE estimator recovers a weighted average of the true effects at different exposure levels. If the true effect is proportional to exposure, the estimate will be attenuated toward zero because "post-treatment" observations include partially-treated recall windows.

Problem 2: Contaminated baselines. More perniciously, the "pre-treatment" period may itself include treatment exposure. In our setting, federal universal meal waivers were active through June 2022, meaning the December 2022 survey—which serves as the reference period for 2023 adopters—captures a recall window (December 2021 through November 2022) that includes 6 months of de facto universal meals nationwide (January through June 2022 under federal waivers). For 2022 state adopters, the recall window additionally includes 4 months of state-specific universal meals (August through November 2022). For 2022 adopters, there is no "pre-treatment" period in the 2022–2024 data at all; they are always treated. For 2023 adopters, the "pre-treatment" reference year (2022) is contaminated by federal waivers. This contamination violates the core DiD assumption that control observations represent the untreated counterfactual.

2.3 Bias from Selection into Treatment

The combination of recall-window mismatch and selection into treatment can produce sign flips—estimated effects opposite in sign to the true causal effect. To see this, decompose the observed outcome:

$$Y_{st} = \underbrace{\alpha_s}_{\text{state FE}} + \underbrace{\delta_t}_{\text{year FE}} + \underbrace{\gamma_s \cdot t}_{\text{state trend}} + \underbrace{\beta \cdot E_{st}}_{\text{true effect}} + \varepsilon_{st} \tag{4}$$

If states that adopt treatment are on *worsening* trajectories ($\gamma_s > 0$ for adopters, meaning food insecurity is rising faster in treated states than in control states), the TWFE estimator picks up both the attenuated treatment effect and the differential trends:

$$\hat{\beta}^{\text{TWFE}} \approx \beta \cdot \bar{E}_{\text{post}} + \underbrace{(\bar{\gamma}_{\text{treated}} - \bar{\gamma}_{\text{control}}) \cdot (\bar{t}_{\text{post}} - \bar{t}_{\text{pre}})}_{\text{selection bias}} \tag{5}$$

When treated states have steeper positive trends ($\bar{\gamma}_{\text{treated}} > \bar{\gamma}_{\text{control}}$)—perhaps because they adopted universal meals *in response to* rising food insecurity—the selection bias term is positive. Combined with severe attenuation of the true negative effect ($\beta \cdot \bar{E}_{\text{post}} \approx 0$), the estimated effect can be positive even though the true effect is negative. This is precisely what the positive TWFE coefficient in Section 5 demonstrates: states may have adopted universal meals because they were experiencing rising food insecurity, not because they had lower baseline rates.

2.4 Formal Conditions for Sign Reversal

Let $\beta < 0$ denote the true full-year treatment effect (treatment reduces food insecurity). Let $\Delta\gamma = \bar{\gamma}_{\text{treated}} - \bar{\gamma}_{\text{control}}$ denote the differential trend (positive if treated states have steeper increases in food insecurity, i.e., are worsening faster or improving slower than controls). Let $\bar{E}$ denote the average exposure intensity in post-treatment observations.

The naive TWFE estimator has the wrong sign (positive) if:

$$\underbrace{\Delta\gamma \cdot (\bar{t}_{\text{post}} - \bar{t}_{\text{pre}})}_{\text{selection component (positive)}} > \underbrace{|\beta| \cdot \bar{E}}_{\text{attenuated treatment effect}} \tag{6}$$

With limited post-periods (small $\bar{t}_{\text{post}} - \bar{t}_{\text{pre}}$), low exposure intensity (small $\bar{E}$), and positive selection ($\Delta\gamma > 0$, states adopting in response to rising food insecurity), sign reversals become likely. Our empirical setting has these features: only 2–3 post-periods, exposure intensity of 0.33 in the first post-year, and treated states that may have adopted precisely because they were experiencing rising food hardship.

2.5 Monte Carlo Evidence

To validate this theoretical framework, I conduct Monte Carlo simulations calibrated to the CPS-FSS setting. The simulation generates:

- 50 states, 5 states adopting treatment in year 3 (August)
- 5 years of survey data with 12-month recall windows

- True treatment effect of −2 percentage points on the annual food insecurity probability when fully exposed (the effect accumulates proportionally with exposure intensity)
- Treated states on *worsening* trajectories (positive trend in food insecurity, reflecting selection into treatment—states adopt universal meals in response to rising food hardship)
- Federal waiver contamination in the "pre-period"

Figure 1 presents results from 500 simulations. The figure shows the distribution of naive TWFE estimates versus exposure-weighted estimates. Despite a true treatment effect of −2 percentage points (treatment reduces food insecurity), *both estimators produce positive coefficients* with means around +6 percentage points. This dramatic sign reversal occurs because the selection effect dominates: treated states were on worsening trajectories (rising food insecurity) that they adopted universal meals to address. The binary treatment indicator captures this differential trend rather than the true causal effect.

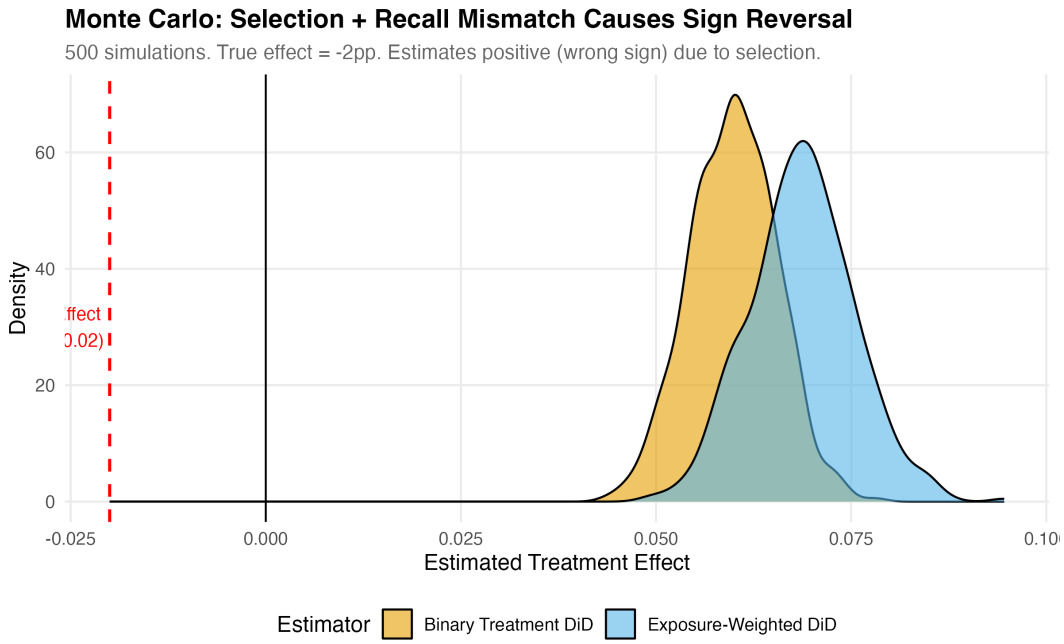

Figure 1: Monte Carlo Evidence: Selection Plus Recall Mismatch Causes Sign Reversal
Notes: Results from 500 Monte Carlo simulations. The vertical red dashed line indicates the true full-year treatment effect (−0.02 or −2 pp). Despite the true negative effect, both estimators produce positive estimates (wrong sign) because treated states were simulated to have worsening food insecurity trends. The selection bias overwhelms the attenuated treatment effect. See text for simulation design details.

These simulation results confirm that recall-window mismatch combined with selection

can produce estimates with the wrong sign. When treated states adopt policy in response to worsening conditions (positive $\Delta\gamma$), the selection bias can overwhelm even a correctly-signed but attenuated treatment effect, producing estimates that suggest treatment *increases* the outcome. The positive coefficient observed in Section 5 is thus consistent with what theory predicts when states selected into treatment based on rising food insecurity—a plausible scenario given that universal meals represent a policy response to perceived hardship.

2.6 Implications for Applied Research

The recall-window mismatch problem arises in many applied settings beyond food security measurement. Examples include:

- **Health surveys:** The National Health Interview Survey (NHIS) asks about conditions "during the past 12 months." Evaluating health policy with annual DiD faces the same exposure-intensity issue.
- **Crime victimization:** The National Crime Victimization Survey (NCVS) asks about victimization over a 6-month reference period. Policy evaluations using NCVS must account for partial exposure.
- **Labor market outcomes:** Annual measures of unemployment duration or job tenure aggregate multiple months of labor market status.
- **Consumption expenditure:** Consumer Expenditure Survey modules ask about spending over various reference periods (2 weeks, 3 months, 12 months).

The general lesson is that researchers must align their treatment coding to the outcome's reference period. When alignment is impossible—as when treatment occurs mid-period—either exposure-weighted estimation or bounds on the partially-identified parameter are required. Standard binary treatment indicators are inappropriate.

3. Institutional Background

3.1 School Meal Programs in the United States

The National School Lunch Program (NSLP), established in 1946, provides reduced-price or free meals to eligible students based on family income. Children from families with incomes below 130% of the federal poverty level qualify for free meals, while those between 130% and 185% qualify for reduced-price meals (currently capped at 40 cents for lunch). The School Breakfast Program (SBP), established in 1966, operates under similar eligibility rules.

Prior to the COVID-19 pandemic, approximately 30 million children received NSLP meals daily, with about 20 million receiving free meals (USDA, 2020). However, even among eligible families, barriers to enrollment persisted. The application process requires documentation of income, which some families may be unable or unwilling to provide. More importantly, the stigma associated with receiving free or reduced-price lunch led many eligible students to skip meals rather than be identified as low-income (Bhatia et al., 2011). This stigma effect has been documented in multiple settings and represents a key inefficiency in means-tested food assistance programs.

The Community Eligibility Provision (CEP), introduced in 2010 and expanded nationally in 2014, allowed high-poverty schools to serve free meals to all students without individual applications if at least 40% of students were "identified" as eligible through other programs (SNAP, TANF, etc.). Research on CEP provides the closest empirical analogue to universal school meals. Schwartz and Rothbart (2020) find that CEP adoption increased academic achievement, particularly for economically disadvantaged students, suggesting that removing barriers to meal access has educational benefits beyond nutrition. However, CEP is limited to high-poverty schools, while the state universal meal policies studied here extend to *all* public schools regardless of poverty concentration.

3.2 The COVID-19 Pandemic and Universal Meals

In response to school closures and economic disruption, the USDA issued waivers in March 2020 allowing all schools to serve free meals to all children regardless of income—effectively implementing temporary universal meals nationwide. These waivers were repeatedly extended through the 2020–21 and 2021–22 school years, covering approximately 30 million students who would not otherwise have qualified for free meals.

When federal waivers expired in June 2022, several states acted to make universal meals permanent through state legislation or appropriations.[1] Table 1 shows the adoption timeline. Five states—California, Maine, Massachusetts, Nevada, and Vermont—adopted universal meals effective August 2022, the start of the 2022–23 school year. Four additional states—Colorado, Michigan, Minnesota, and New Mexico—followed in August 2023 for the 2023–24 school year.

[1]Adoption dates are based on the USDA Economic Research Service classification used in Rabbitt et al. (2024) and verified against the Food Research & Action Center's state legislation tracker. Nevada funded universal meals for the 2022–23 school year using $75 million in American Rescue Plan funds approved by the Interim Finance Committee; this is treated as a 2022 adoption per the USDA classification.

Table 1: Universal Free School Meals Adoption by State

State	Effective Date	School Year
2022 Adopters (5 states)		
California	August 2022	2022–23
Maine	August 2022	2022–23
Massachusetts	August 2022	2022–23
Nevada	August 2022	2022–23
Vermont	August 2022	2022–23
2023 Adopters (4 states)		
Colorado	August 2023	2023–24
Michigan	August 2023	2023–24
Minnesota	August 2023	2023–24
New Mexico	August 2023	2023–24

Notes: Table shows states that adopted universal free school meals for all K–12 public school students through state legislation or appropriations, effective as of the dates shown. Adoption dates verified against USDA ERS classification (Rabbitt et al., 2024) and the Food Research & Action Center state legislation tracker. Nevada extended its program through the 2023–24 school year using additional ARPA funds; the program was discontinued for 2024–25. Washington DC, despite having free breakfast for all students, has not enacted universal free lunch and is classified as a control unit. The 41 remaining states plus DC (42 control units total) serve as potential controls, yielding 51 state/DC units in total.

The adopting states are not randomly selected. California and Massachusetts are large, wealthy states with relatively strong safety nets. Vermont has among the lowest food insecurity rates in the nation. Nevada and New Mexico, conversely, have relatively high food insecurity. This heterogeneity in adopter characteristics raises concerns about selection into treatment that would violate the parallel trends assumption even with adequate pre-treatment data.

3.3 The Resource Reallocation Hypothesis

The direct effect of school meals on child nutrition is well-established in the literature (Hinrichs, 2010; Gundersen and Ziliak, 2015). Universal meals may additionally affect household-level food security through at least three channels.

First, for families who previously paid reduced-price copays or full price for school meals, universal meals represent a direct financial transfer. A family paying $0.40 per lunch and $0.30 per breakfast for two children saves approximately $250 per school year—modest but

non-trivial for low-income households.

Second, for families who were already eligible for free meals but did not participate due to stigma or administrative barriers, universal meals may increase uptake. If the marginal non-participant is food insecure, bringing them into the program should reduce measured food insecurity.

Third, even for families already receiving free meals, universal provision may reduce "churn"—the phenomenon of losing eligibility due to income fluctuations or paperwork failures—ensuring more consistent access to meals throughout the school year.

Recent evidence from USDA researchers provides suggestive support for the resource reallocation hypothesis. Rabbitt et al. (2024) find that states with universal meal policies experienced 1–2 percentage point reductions in child food insufficiency during the 2022–23 school year using the Household Pulse Survey. However, the Household Pulse Survey uses a 7-day recall window, which aligns more cleanly with treatment timing than the 12-month recall window in CPS-FSS used in this paper.

4. Data and Empirical Strategy

4.1 Data

The Current Population Survey Food Security Supplement (CPS-FSS) is the primary data source for official statistics on food insecurity in the United States. Conducted annually in December by the Census Bureau for the USDA Economic Research Service, the CPS-FSS is a nationally representative household survey that appends a food security module to the standard CPS monthly labor force survey. The CPS-FSS has been conducted annually since 1995, providing a long time series for food security research, though my analysis focuses on the 2022–2024 window due to the recency of state universal meal adoptions.

The food security module asks about conditions and behaviors during the *prior 12 months* that indicate food insecurity, such as being unable to afford balanced meals, cutting meal sizes due to insufficient money, or going hungry because there was not enough food (USDA, 2000). Responses are aggregated into an 18-item scale, and households are classified into three categories: food secure (no or minimal evidence of food access problems), low food security (reduced quality, variety, or desirability of diet), and very low food security (disrupted eating patterns and reduced food intake). I construct a binary indicator equal to one for households with low or very low food security.

The measurement properties of the CPS-FSS food security module are well-established. The 18-item scale has high internal consistency (Cronbach's alpha > 0.85) and has been validated against multiple external criteria including food expenditures, dietary intake, and

program participation (Bound, Brown & Mathiowetz, 2001; Tourangeau, Rips & Rasinski, 2000). However, the 12-month recall window introduces challenges for policy evaluation that have received less attention in the literature. Survey response theory suggests that respondents may anchor on salient recent events when answering retrospective questions, potentially overweighting conditions near the survey date relative to conditions early in the recall window (Tourangeau, Rips & Rasinski, 2000). This "recency bias" could attenuate estimated effects of policies that began mid-period.

The sample is restricted to households with at least one child aged 5–17 (school age), as these households are directly affected by universal meal policies. Data are available for survey years 2022–2024. Critically, the December 2022 survey asks about food security from approximately December 2021 through November 2022, the December 2023 survey covers December 2022 through November 2023, and the December 2024 survey covers December 2023 through November 2024. This 12-month recall window creates a fundamental measurement challenge discussed in detail in Section 2 and the empirical strategy below.

For the triple-difference specification, I also include households without school-age children as a within-state-year control group. These households should not be directly affected by universal school meal policies, though they may be indirectly affected through local economic conditions or general equilibrium effects on food prices. The DDD sample uses the same state restriction as the TWFE/C-S regressions (2023 adopters plus never-treated, i.e., 46 states/DC) but expands to include all household types in those states.

Table 2 presents summary statistics for the full descriptive sample. This sample includes all 9 states that eventually adopt universal meals (the "ever-treated" group) compared to 42 never-treated states/DC. Treatment states have slightly lower average food insecurity (12.7%) than control states (13.5%), consistent with selection—states that adopt progressive food policies tend to have stronger safety nets overall.

Note that the *estimation samples* used in regression analyses differ from this descriptive sample. Because 2022 adopters are always treated in the 2022–2024 data and contribute no within-state variation, the TWFE and Callaway-Sant'Anna regressions restrict to **2023 adopters (4 states: CO, MI, MN, NM) plus 42 never-treated states/DC**, yielding N = 23,489 for households with school-age children. The triple-difference specification uses the same state restriction but expands to include households without school-age children (N = 107,871). See table notes for sample details.

Table 2: Summary Statistics: Full Descriptive Sample (All Ever-Treated States)

	Control States	Treatment States	Difference
Food insecurity rate (%)	13.5	12.7	−0.8
Very low food security (%)	4.2	3.7	−0.5
Observations (household-years)	21,800	5,714	
States/DC	42	9	

Notes: **Full descriptive sample** includes households with at least one child aged 5–17 from the CPS Food Security Supplement 2022–2024. Treatment states include all 9 ever-adopters: CA, CO, MA, ME, MI, MN, NM, NV, VT. Control states include 42 never-treated units (41 states plus DC). Statistics are pooled across survey years. **Regression samples differ**: TWFE/C-S use only 2023 adopters (4 states) + never-treated; DDD adds households without children. See regression table notes for exact sample definitions.

4.2 Empirical Strategy

4.2.1 The Recall-Window Mismatch Problem

A fundamental challenge in this analysis is the mismatch between treatment timing and outcome measurement. The CPS-FSS measures food security over a rolling 12-month recall window, while treatment—adoption of universal school meals—occurs at a discrete point in time (August of the adoption year).

To formalize this problem, let Y_{ist} denote the food security status of household i in state s as measured in survey year t. This outcome reflects conditions over the 12-month period from approximately December of year $t-1$ through November of year t. Let $D_{s\tau}$ be an indicator for whether state s has universal meals in effect during month τ. For states adopting in August of year g, we have $D_{s\tau} = 1$ for $\tau \geq \text{August}(g)$ and $D_{s\tau} = 0$ otherwise.

The true exposure to treatment for a household surveyed in December of year t is:

$$\text{Exposure}_{st} = \frac{1}{12} \sum_{\tau \in \text{Dec}(t-1) \text{ to Nov}(t)} D_{s\tau} \tag{7}$$

For an August g adopter surveyed in December of year g, this exposure equals approximately $4/12 = 0.33$—only 4 months (August through November) of the 12-month recall window fall after policy adoption.[2] For a survey in December of year $g+1$, exposure equals $12/12 = 1.0$ (the full recall window from December g through November $g+1$ falls after the

[2]Schools typically begin in late August, so the August effective date captures approximately the full school year. For simplicity, I treat August as a fully treated month.

August g adoption date).

The standard DiD treatment indicator $\text{Treated}_{st} = \mathbb{I}[t \geq g]$ ignores this exposure heterogeneity. It codes both the December g survey (33% exposure) and December $g+1$ survey (100% exposure) as "treated = 1," while the December $g-1$ survey is coded as "treated = 0" (0% exposure). This creates two problems:

First, the "pre-treatment" period (survey year $g-1$) may still include some treated months if the recall window spans the adoption date. For August 2022 adopters, the December 2022 survey's recall window (December 2021 through November 2022) includes 4 treated months (August through November 2022). More importantly, federal universal-meal waivers were in effect through June 2022, meaning 6 months of the December 2022 recall window (January through June 2022) had de facto universal meals *nationwide*. This blurs any distinction between treatment and control for that survey year.

Second, the "post-treatment" periods for staggered adopters have different effective exposure intensities that are not captured by a binary treatment indicator. Comparing December 2023 outcomes for August 2023 adopters (33% exposure) to August 2022 adopters (100% exposure) conflates timing and treatment intensity.

4.2.2 A Misspecified Estimation (For Illustration)

To demonstrate the consequences of ignoring the recall-window mismatch, I estimate a standard two-way fixed effects model:

$$Y_{ist} = \beta \cdot \text{Treated}_{st} + \gamma_s + \delta_t + \varepsilon_{ist} \tag{8}$$

where Treated_{st} equals one if state s has adopted universal meals by survey year t, γ_s are state fixed effects, and δ_t are year fixed effects. Standard errors are clustered at the state level.

Following Goodman-Bacon (2021), I restrict the sample to 2023 adopters plus never-treated states, excluding 2022 adopters who are always-treated in the 2022–2024 sample window and thus contribute no within-state variation under state fixed effects. This avoids the "already-treated-as-controls" problem identified in the staggered DiD literature (Borusyak, Jaravel & Spiess, 2024; Sun and Abraham, 2021).

This specification is **intentionally misspecified** and does not estimate a well-defined treatment effect. The purpose of reporting it is to illustrate what goes wrong when standard DiD machinery is applied to data with timing mismatch—not as a credible policy estimate. A correct specification would require either (1) an exposure-intensity measure replacing the binary treatment indicator, or (2) data where the outcome reference period aligns with

treatment timing. Neither is feasible with CPS-FSS 2022–2024.

4.2.3 Callaway-Sant'Anna Estimator

I also implement the Callaway and Sant'Anna (2021) group-time ATT estimator, which is robust to treatment-effect heterogeneity across adoption cohorts and time. Using 2022 as the reference period for 2023 adopters and never-treated states as controls, I estimate group-time average treatment effects $ATT(g = 2023, t)$ for $t \in \{2023, 2024\}$. These are then aggregated to an overall ATT. Standard errors are computed via multiplier bootstrap with 1,000 iterations.

While the Callaway-Sant'Anna estimator addresses heterogeneous treatment effects, it does not resolve the recall-window mismatch. The estimated $ATT(2023, 2023)$ still compares a 33%-exposed outcome to a 0%-exposed reference period—and that reference period (2022) is contaminated by federal waivers.

4.2.4 Triple-Difference Specification

A more informative specification exploits within-state, within-year variation in treatment exposure. Households *without* school-age children in the same state and year provide a control group that should not be directly exposed to universal school meals policy. The triple-difference (DDD) specification with state×year fixed effects is:

$$Y_{ist} = \alpha_{st} + \beta_1 \cdot \text{Child}_i + \beta_2 \cdot \text{Child}_i \times \text{Treated}_{st} + \varepsilon_{ist} \tag{9}$$

where α_{st} are state×year fixed effects that absorb all state-year-specific variation (including the main effect of treatment), and Child_i indicates whether household i has at least one child aged 5–17. The coefficient β_2 captures the differential effect of treatment on households with versus without school-age children, identified by within-state-year comparisons. If universal meals reduce food insecurity specifically for households with school-age children relative to childless households in the same state-year, we should observe $\beta_2 < 0$.

4.2.5 Exposure-Intensity Estimation (Theoretical Alternative)

A theoretically motivated alternative to binary treatment coding is exposure-intensity estimation. Rather than coding $\text{Treated}_{st} \in \{0, 1\}$, one could construct:

$$\text{Exposure}_{st} = \frac{\text{Months of policy exposure in recall window}}{12} \tag{10}$$

For August adopters surveyed in December of the adoption year, Exposure $= 4/12 \approx 0.33$

(August through November). For the same states surveyed one year later, Exposure = 12/12 = 1. Control states have Exposure = 0 throughout.

The exposure-weighted specification would be:

$$Y_{ist} = \beta \cdot \text{Exposure}_{st} + \gamma_s + \delta_t + \varepsilon_{ist} \tag{11}$$

Under the assumption that the true treatment effect scales linearly with exposure (a reasonable approximation for continuous treatments like monthly meal provision), β estimates the effect of moving from zero to full-year exposure. This estimand is well-defined even when treatment begins mid-period.

However, exposure-intensity estimation does not resolve the contaminated baseline problem. If the "pre-period" recall window includes policy exposure (as it does here, due to federal waivers through June 2022), the exposure measure for control observations is misspecified. I do not implement exposure-intensity estimation in this paper because the baseline contamination renders even this approach invalid for our setting. Future work with longer time series could extend backward to establish clean pre-waiver baselines and then apply exposure-intensity coding to the post-waiver differential adoption period.

4.2.6 Inference with Few Treated Clusters

With only 4 treated states in the 2023 cohort (the sample with identifying variation), standard cluster-robust standard errors may be unreliable (MacKinnon, Nielsen & Webb, 2022; Cameron and Miller, 2015; Bertrand, Duflo & Mullainathan, 2004). I supplement standard inference with randomization inference (Ferman and Pinto, 2019; Conley and Taber, 2011), which permutes treatment assignment across states and computes the distribution of treatment effects under the null hypothesis of no effect. The proportion of permuted effects exceeding the observed effect provides a finite-sample valid p-value. Wild cluster bootstrap was attempted but failed due to collinearity induced by Rademacher weight assignments, a common issue with very few treated clusters.

5. Results

5.1 Descriptive Evidence

Figure 2 plots food insecurity rates for ever-treated versus never-treated states over 2022–2024. Ever-treated states (those that adopted universal meals in either 2022 or 2023) have consistently lower food insecurity than never-treated states across all years. However, both groups show relatively stable trends with no clear visual divergence after treatment.

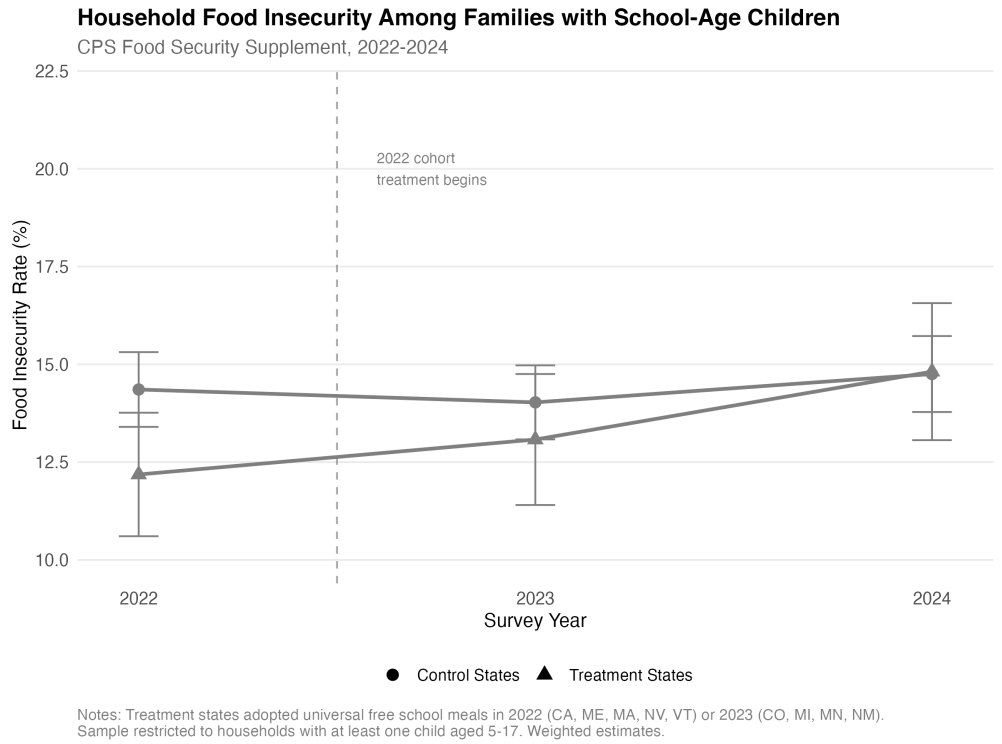

Figure 2: Food Insecurity Trends: Ever-Treated vs. Never-Treated States

Notes: Figure shows weighted mean food insecurity rates for households with at least one child aged 5–17. "Treatment States" are those that adopted universal free school meals at any point (CA, ME, MA, NV, VT in 2022; CO, MI, MN, NM in 2023)—this is an "ever-treated" grouping that combines both cohorts. Note that in 2022, the 2023 adopters are not yet treated, though all states were under federal universal-meal waivers through June 2022. Error bars show 95% confidence intervals.

The level difference—treatment states have approximately 1–2 percentage points lower food insecurity throughout the sample—is consistent with selection. States that adopt progressive food policies have different baseline characteristics: stronger safety nets, higher median incomes, different demographic compositions. This is a "levels" comparison, not a causal estimate, and underscores the importance of the parallel trends assumption for any DiD interpretation.

5.2 TWFE Estimates (Misspecified)

Table 3 presents TWFE estimates that **should not be interpreted as treatment effects**. These results demonstrate what happens when standard DiD methods are applied to data with severe timing mismatch.

Table 3: TWFE Estimates (Demonstration of Design Failure)

	(1) Food Insecure	(2) Very Low FS
Treated	0.047**	0.021**
	(0.020)	(0.010)
	[0.009, 0.085]	[0.002, 0.040]
State FE	Yes	Yes
Year FE	Yes	Yes
Observations	23,489	23,489
Clusters (states)	46	46
Treated states	4	4
Mean (never-treated)	0.135	0.042

Notes: Restricted sample: 2023 adopters (CO, MI, MN, NM) plus 42 never-treated states/DC. 2022 adopters excluded. Standard errors clustered at state level in parentheses. 95% confidence intervals in brackets. Weighted by CPS household weights. **Warning:** These estimates are **not interpretable as treatment effects**. The treatment indicator is misaligned with the 12-month recall window outcome. For 2023 adopters, "Treated=1" in 2023 corresponds to only ∼4 months of policy exposure (Aug–Nov) in the 12-month outcome window.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The coefficient on food insecurity is 0.047 (SE = 0.020, 95% CI: [0.009, 0.085]). This is statistically significant at the 5% level but substantively implausible. Universal meals provide free food; they cannot causally increase food hardship through any plausible mechanism. The positive sign is a diagnostic signal that the specification is capturing spurious variation rather than causal effects.

5.3 Callaway-Sant'Anna Estimates

Table 4 presents Callaway-Sant'Anna group-time ATT estimates for the 2023 cohort using never-treated states as controls and 2022 as the reference period.

Table 4: Callaway-Sant'Anna Estimates (2023 Cohort Only)

Group	Time	ATT(g,t)	SE
2023	2023	0.046*	(0.023)
2023	2024	0.057*	(0.024)
Aggregated ATT		0.052**	(0.024)
95% CI		[0.005, 0.099]	
Household observations (N)		23,489	
State-year cells		138	
States (4 treated, 42 control)		46	

Notes: Callaway-Sant'Anna estimator with never-treated controls. Estimation proceeds at the state-year level using weighted means; household N reflects the underlying microdata. Standard errors via multiplier bootstrap (1,000 iterations). * indicates 95% uniform confidence band excludes zero. **Warning:** These estimates inherit all timing/recall issues from the TWFE analysis. The reference year (2022) is contaminated by federal waivers, and "post-treatment" years have partial exposure.

The aggregated ATT is 0.052 (SE = 0.024, 95% CI: [0.005, 0.099]), marginally significant. The group-time estimates show similar positive coefficients in both post-treatment years. However, these estimates are subject to the same recall-window mismatch as the TWFE specification. The Callaway-Sant'Anna estimator addresses treatment-effect heterogeneity across cohorts but cannot resolve measurement misalignment between treatment and outcome timing.

5.4 Triple-Difference Results

Table 5 presents the triple-difference specification comparing households with and without school-age children.

Table 5: Triple-Difference Estimates (State×Year Fixed Effects)

	Food Insecure
Has Children (5–17)	0.053***
	(0.005)
Has Children × Treated	−0.008
	(0.013)
	[−0.034, 0.018]
State×Year FE	Yes
Observations	107,871
State-year cells	138

Notes: Full sample including households with and without school-age children. 2023 adopters + never-treated only. State×year fixed effects absorb all state-year variation including the main treatment effect. Standard errors clustered at state level. 95% CI in brackets for the interaction term. The DDD coefficient is identified by within-state-year comparisons: how much more (or less) did food insecurity change for households with children relative to households without children in treatment vs. control states after adoption?

The triple-difference estimate is −0.008 (SE = 0.013, 95% CI: [−0.034, 0.018]). This is substantively small and statistically indistinguishable from zero. The point estimate is negative, as theory would predict if universal meals reduce food insecurity for households with school-age children relative to households without, but the effect is precisely estimated to be near zero.

With state×year fixed effects, identification comes entirely from within-state-year variation. The DDD compares how the gap in food insecurity between households with and without school-age children changes across treatment and control states after adoption. Under the assumption that this gap would have evolved similarly in treatment and control states absent the policy, the coefficient β_2 estimates the causal effect of universal meals on households with children relative to the counterfactual trend. The near-zero estimate suggests either that

universal meals have no detectable effect on household food insecurity, or that the within-state-year comparison lacks power to detect meaningful effects given the short exposure window.

5.5 Robustness: Randomization Inference

Table 6 summarizes inference from alternative procedures designed for settings with few treated clusters.

Table 6: Summary of Main Estimates

Specification	Estimate	p-value	95% CI
TWFE (cluster-robust SE)	0.047	0.017	[0.009, 0.085]
TWFE (randomization inference)	0.047	0.015	—
DDD (state×year FE)	−0.008	0.547	[−0.034, 0.018]

Notes: TWFE uses state-clustered standard errors (46 clusters). Randomization inference permutes treatment across 999 placebo state assignments. DDD uses state×year fixed effects to absorb all state-year variation and compares households with vs. without school-age children within state-years.

Randomization inference yields a p-value of 0.015, similar to the cluster-robust p-value of 0.017. This suggests the positive TWFE coefficient is unlikely under random assignment of treatment to states. However, statistical significance does not imply causal interpretation when the estimand is poorly defined. The triple-difference specification, which provides a cleaner comparison, finds no significant effect ($p = 0.547$).

Figure 3 shows the permutation distribution from randomization inference. The observed effect of 0.047 lies in the upper tail, but the distribution is not centered at zero—reflecting the substantial variation in state-level food insecurity trends unrelated to treatment.

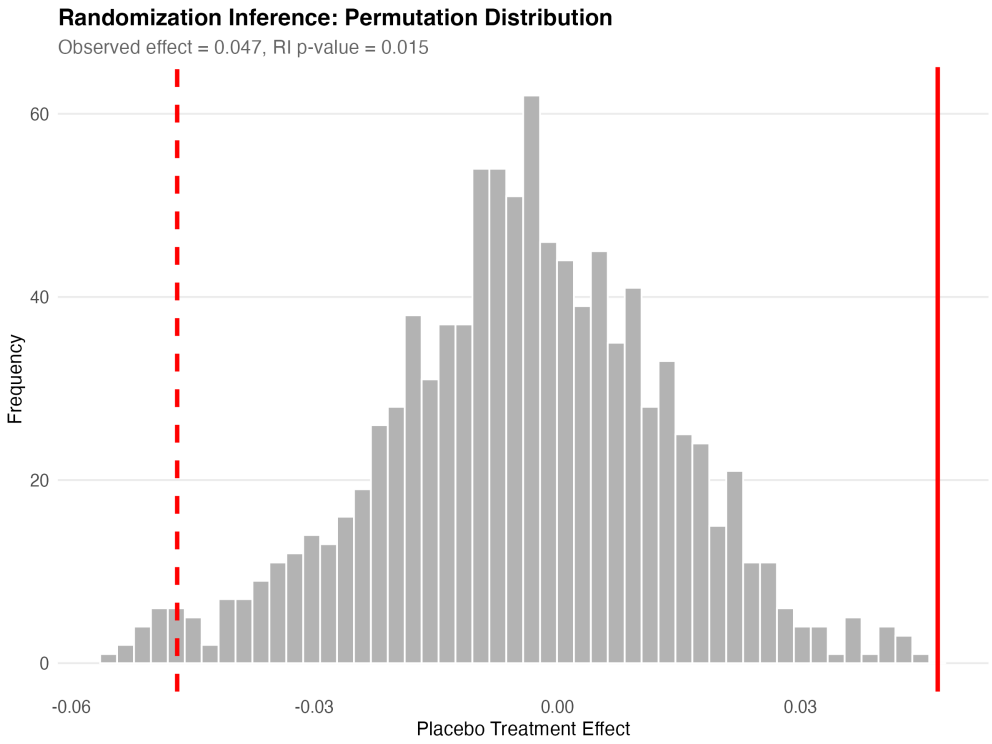

Figure 3: Randomization Inference: Permutation Distribution

Notes: Histogram shows distribution of placebo treatment effects from 999 permutations that randomly assign treatment to 4 states. Red vertical line indicates observed effect (0.047). Dashed red line indicates the symmetric negative value. RI p-value = 0.015.

6. Discussion: Why These Estimates Are Not Causal

The positive TWFE and Callaway-Sant'Anna estimates should not be interpreted as evidence that universal school meals increase food insecurity. The counterintuitive sign is a diagnostic signal of design failure, not a substantive finding. This section discusses the specific identification challenges and what would be required for credible inference.

6.1 Recall-Window Mismatch Invalidates the Estimand

The most fundamental problem is that the treatment indicator does not correspond to the outcome's reference period. When the CPS-FSS asks respondents in December 2023 about food security "during the last 12 months," the response integrates conditions from December 2022 through November 2023. For states adopting universal meals in August 2023, this 12-month window includes 8 months without the policy (December 2022 through July 2023) and only 4 months with the policy (August through November 2023)—approximately 33% exposure.

Coding such a household as "treated" conflates a mixed-exposure outcome with a discrete treatment indicator. The coefficient β does not estimate a well-defined causal parameter—it is a weighted average of effects at different exposure intensities, contaminated by pre-treatment conditions included in the outcome measure.

This problem is not unique to school meals or food security. Any DiD analysis using survey data with rolling recall windows faces analogous challenges. Health surveys asking about conditions "in the past 12 months" (e.g., disability, chronic illness episodes), crime victimization surveys, and consumption expenditure modules with annual reference periods all require careful attention to the alignment between treatment timing and outcome measurement.

6.2 Absence of Pre-Treatment Data Precludes Parallel Trends Assessment

With my CPS-FSS sample covering only 2022–2024, there are zero pre-treatment observations for states adopting in August 2022. For 2023 adopters, only the December 2022 survey provides a "pre-treatment" data point—but as discussed above, this survey's recall window includes the final months of federal universal-meal waivers (January through June 2022), when all states had de facto universal meals.

The parallel trends assumption—that treatment and control states would have followed the same trajectory absent the policy—cannot be assessed, let alone validated, without pre-treatment data. The recent DiD literature has emphasized that failing to reject a pre-trends test does not establish parallel trends (Roth, 2022), but here we cannot even conduct such a test.

The finding that treatment states have *lower* food insecurity throughout the sample period (see Figure 2) is consistent with selection: states with stronger safety nets and lower baseline food insecurity are more likely to adopt progressive food policies. This level difference raises concerns that treatment and control states may be on fundamentally different trajectories even absent the policy.

6.3 Coincident Policy Changes Confound Treatment Timing

The study period 2022–2024 coincides with major federal policy changes in food assistance that differentially affected states. SNAP Emergency Allotments, which increased monthly benefits by approximately $200 for the average participating household, ended in March 2023. Pandemic EBT (P-EBT), which provided food benefits to replace school meals during COVID-19 closures, was phased out throughout 2022–2023. Food price inflation reached 40-year highs in 2022 before moderating in 2023–2024.

If treatment states experienced different trajectories of economic recovery, had different SNAP participation rates, or were differentially affected by inflation, these factors would confound the DiD estimates. The triple-difference specification attempts to address state-level confounds by comparing households with and without children within the same state-year, but it cannot address confounds that differentially affect households with school-age children across treatment and control states.

6.4 Few Treated Clusters Limit Inference

With only 9 treatment states (and effectively 4 contributing identifying variation in the 2023 cohort), inference is fragile. Standard asymptotic approximations for cluster-robust standard errors require a large number of clusters (MacKinnon, Nielsen & Webb, 2022). Alternative methods—wild cluster bootstrap, randomization inference—provide finite-sample valid inference but do not address the fundamental identification problems.

The randomization inference p-value of 0.015 indicates that observing a coefficient as large as 0.047 would be unlikely if treatment were randomly assigned across states. However, treatment was not randomly assigned, and the statistical significance reflects the combination of true state-level differences in food insecurity trends and the mechanical correlation induced by recall-window mismatch.

6.5 What Would Credible Identification Require?

A credible evaluation of universal school meals' effects on household food insecurity would require several elements not present in this analysis.

First, multiple years of pre-treatment data—ideally 5 or more years—would enable assessment of parallel trends using event-study specifications and sensitivity analyses such as HonestDiD bounds (Rambachan and Roth, 2023). With CPS-FSS data from 2015–2024 (available through IPUMS), researchers could construct event studies for both the 2022 and 2023 adoption cohorts with adequate pre-periods. The longer time series would also allow researchers to observe how treatment and control states responded to previous shocks (e.g., the 2008–2009 recession, SNAP benefit changes) and assess the plausibility of parallel trends during the study period.

Second, an exposure-intensity measure would properly account for the recall-window mismatch. Rather than a binary treatment indicator, the regressor should measure the fraction of months in the recall window during which the policy was in effect. This requires knowing both the exact policy effective date (available) and the exact recall window dates for each respondent (approximated as December $t-1$ through November t). Section 4.2.5

outlines this approach, though it does not resolve baseline contamination.

Third, explicit controls for coincident policy changes—SNAP EA termination dates, P-EBT disbursements, state unemployment rates, food price inflation—would address the most obvious confounders. Alternatively, a triple-difference design using households without school-age children as within-state controls (as implemented here) provides one approach, though it cannot address confounds specific to households with children.

Fourth, appropriate inference for small numbers of treated clusters—using wild cluster bootstrap, randomization inference, or Conley-Taber confidence intervals (Conley and Taber, 2011)—would provide reliable p-values and confidence intervals.

Fifth, the federal waiver period (2020–2022) should be modeled explicitly rather than ignored. One approach would treat the waiver as a nationwide temporary treatment and analyze the post-waiver period as differential continuation of universal meals. This requires data extending back to 2019 or earlier to establish pre-waiver baselines.

6.6 Alternative Data and Identification Strategies

Several alternative approaches could provide more credible evidence on universal meal effects, though each has limitations.

Household Pulse Survey. The Census Bureau's Household Pulse Survey measures food insufficiency over a 7-day reference period, providing much closer alignment between treatment timing and outcome measurement. Rabbitt et al. (2024) use Household Pulse to find suggestive evidence of reduced child food insufficiency in universal meal states. However, the Household Pulse has smaller sample sizes and higher nonresponse rates than CPS-FSS, and its brief reference period may miss effects that accumulate over longer horizons.

Administrative data. School district administrative data on meal participation, meal counts, and potentially matched student information could provide a first-stage (did universal meals increase participation?) and potentially student-level outcomes. However, linking to household food security is challenging, and administrative data typically lacks information on non-participating households.

Border-county designs. If CPS microdata allows county identification (restricted access), researchers could compare households in counties along state borders where one state adopted universal meals and the neighboring state did not. This design requires sufficient sample sizes in border counties and faces its own threats from cross-border spillovers.

Regression discontinuity at eligibility thresholds. Prior to universal meals, eligibility for free and reduced-price meals varied discontinuously at 130% and 185% of the federal poverty level. Comparing households just above and below these thresholds provides a different estimand (effect of meal subsidies at the margin) but could inform the resource

reallocation mechanism (Grogger, 2019).

6.7 Broader Implications for Survey-Based Policy Evaluation

The challenges documented in this paper extend beyond school meal policy to any setting where survey outcomes use rolling recall windows. The core problem—that binary treatment coding conflates different exposure intensities and can contaminate baselines—is generic to many policy evaluation contexts.

The literature on difference-in-differences has made remarkable progress on issues like heterogeneous treatment effects across adoption cohorts (Goodman-Bacon, 2021; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; de Chaisemartin and D'Haultfoeuille, 2020; Borusyak, Jaravel & Spiess, 2024) and sensitivity to parallel trends violations (Rambachan and Roth, 2023; Roth, 2022). However, the measurement alignment problem has received comparatively less attention. This paper highlights that even with the most sophisticated estimators, credible inference requires that the outcome measure be coherently defined relative to treatment timing.

Applied researchers should routinely ask: (1) What is the reference period for my outcome measure? (2) Does my treatment coding align with this reference period? (3) If not, what fraction of the outcome's reference period is actually "treated"? (4) Is my "pre-period" contaminated by partial treatment or related policies?

When alignment is imperfect, researchers have several options: construct exposure-intensity measures, use outcomes with shorter reference periods, restrict the sample to observations with clean exposure (either fully treated or fully untreated recall windows), or explicitly bound the partially-identified treatment effect. Simply ignoring the mismatch—as a naive application of standard DiD would do—can produce misleading results including sign reversals.

7. Conclusion

This paper attempted to evaluate whether universal free school meals reduce household-level food insecurity through a resource reallocation mechanism. The answer is: we cannot know from this analysis. The combination of recall-window mismatch, absence of pre-treatment data, coincident policy shocks, and selection into treatment makes credible causal inference impossible with the available data structure.

The positive TWFE estimate (4.7 percentage points more food insecurity in treatment states) is almost certainly an artifact of identification failure rather than a true causal effect. Universal meals provide free food to children; no plausible mechanism would cause

them to increase food insecurity. The triple-difference estimate (-0.8 percentage points, not significant) is more credible and suggests approximately zero effect when state-year confounds are absorbed via state×year fixed effects—but this null finding could also reflect insufficient exposure intensity in the short post-treatment window or the triple-diff's reliance on parallel trends between households with and without children within treatment states.

The broader methodological contribution of this paper is to formalize and illustrate the recall-window mismatch problem for difference-in-differences designs. Section 2 develops a theoretical framework showing how binary treatment coding, when applied to outcomes with rolling recall windows, produces attenuated and potentially sign-flipped estimates. Monte Carlo simulations calibrated to the CPS-FSS setting confirm that even large true treatment effects can be obscured or reversed when exposure intensity varies within "post-treatment" observations and when selection into treatment is present.

The implications extend beyond food security and school meal policy. Any applied researcher using survey data with retrospective reference periods—health surveys, crime victimization, consumption expenditure, labor market measures—faces analogous challenges. The key lessons are:

First, **check reference period alignment**. Before estimating a DiD model, verify that the outcome's reference period aligns with treatment timing. When treatment occurs mid-period, standard binary coding is inappropriate.

Second, **construct exposure-intensity measures**. When alignment is imperfect, replace binary treatment indicators with the fraction of the reference period actually exposed to treatment. This provides a well-defined estimand even under partial exposure.

Third, **verify baseline cleanliness**. Even with exposure-intensity coding, DiD requires that the "pre-period" represents the untreated counterfactual. If related policies (e.g., federal waivers) contaminate the baseline, identification fails regardless of sophisticated estimators.

Fourth, **use multiple pre-treatment periods**. With adequate pre-treatment data, researchers can assess parallel trends and apply sensitivity analyses like HonestDiD bounds. Without such data—as in this paper's 2022–2024 window—even perfectly aligned treatment coding cannot rescue a fundamentally under-identified design.

Future work on universal school meals should extend the CPS-FSS sample back to 2015 or earlier through IPUMS, explicitly model the federal waiver period as a nationwide regime shift, and construct proper exposure-intensity measures. Alternative identification strategies—Household Pulse Survey with its 7-day reference period, administrative data on meal participation, border-county designs, or regression discontinuity at income eligibility thresholds—could provide complementary evidence. Until then, the suggestive findings from Rabbitt et al. (2024) using Household Pulse data provide the best available, if preliminary,

evidence on the household-level effects of universal school meal policies.

The cautionary tale documented here should not discourage policy evaluation. Rather, it should encourage researchers to invest in proper research design *before* attempting estimation. The recall-window mismatch problem is detectable at the design stage; researchers who check reference period alignment before running regressions can avoid the misleading results demonstrated in this paper. Applied econometrics has made tremendous progress on inference and heterogeneous effects; what remains underappreciated is that credible identification starts with measurement.

References

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

Bhattacharya, J., Currie, J., and Haider, S. (2006). Breakfast of champions? The School Breakfast Program and the nutrition of children and families. *Journal of Human Resources*, 41(3), 445–466.

Bhatia, R., Jones, P., and Reicker, Z. (2011). Competitive foods, discrimination, and participation in the National School Lunch Program. *American Journal of Public Health*, 101(8), 1380–1386.

Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement error in survey data. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Vol. 5 (pp. 3705–3843). Elsevier.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6), 3253–3295.

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

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.

Coleman-Jensen, A., Rabbitt, M. P., Gregory, C. A., and Singh, A. (2024). Household food security in the United States in 2023. USDA Economic Research Service Report.

Conley, T. G. and Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *Review of Economics and Statistics*, 93(1), 113–125.

de Chaisemartin, C. and D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9), 2964–2996.

Ferman, B. and Pinto, C. (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *Review of Economics and Statistics*, 101(3), 452–467.

Frisvold, D. E. (2015). Nutrition and cognitive achievement: An evaluation of the School Breakfast Program. *Journal of Public Economics*, 124, 91–104.

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

Gordon, A., Briefel, R., Collins, A., Rowe, G., and Klerman, J. A. (2007). School nutrition dietary assessment study-III: Volume I. USDA Food and Nutrition Service.

Grogger, J. (2019). Are current eligibility guidelines for free and reduced-price meals effective? *Journal of Policy Analysis and Management*, 38(4), 893–919.

Gundersen, C. and Ziliak, J. P. (2015). Food insecurity and health outcomes. *Health Affairs*, 34(11), 1830–1839.

Hinrichs, P. (2010). The effects of the National School Lunch Program on education and health. *Journal of Policy Analysis and Management*, 29(3), 479–505.

Hoynes, H. W., Schanzenbach, D. W., and Almond, D. (2016). Long-run impacts of childhood access to the safety net. *American Economic Review*, 106(4), 903–934.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2022). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*.

Rabbitt, M. P., Reed-Jones, M., Hales, L. J., and Burke, M. P. (2024). State universal free school meal policies reduced food insufficiency among children. *USDA ERS Amber Waves*.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5), 2555–2591.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3), 305–322.

Schwartz, A. E. and Rothbart, M. W. (2020). Let them eat lunch: The impact of universal free meals on student performance. *Journal of Policy Analysis and Management*, 39(2), 376–410.

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

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Bickel, G., Nord, M., Price, C., Hamilton, W., and Cook, J. (2000). Guide to measuring household food security. USDA Food and Nutrition Service.

USDA Food and Nutrition Service. (2020). National School Lunch Program participation and meals served.

A. Appendix: Additional Figures

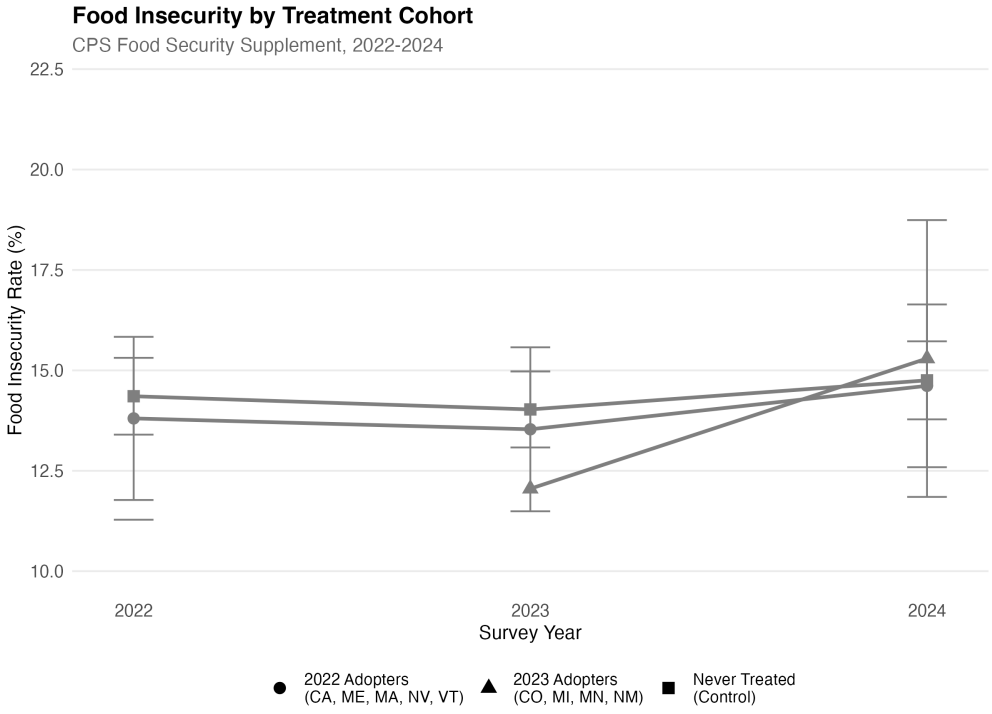

Figure 4: Food Insecurity by Treatment Cohort

Notes: Figure shows weighted mean food insecurity rates by treatment cohort. 2022 Adopters: CA, ME, MA, NV, VT. 2023 Adopters: CO, MI, MN, NM. Error bars show 95% confidence intervals.

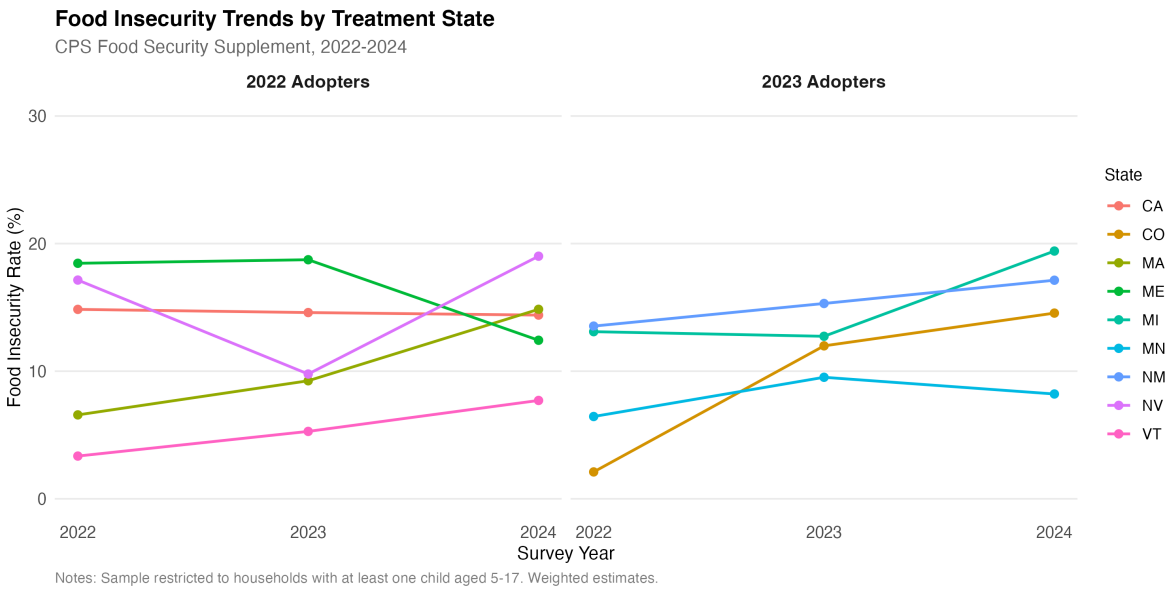

Figure 5: Food Insecurity Trends by Individual Treatment State

Notes: Figure shows weighted mean food insecurity rates for each treatment state separately. Substantial heterogeneity is evident, with Vermont (VT) showing much lower food insecurity than Nevada (NV) or New Mexico (NM).

Acknowledgements

This paper was autonomously generated as part of the Autonomous Policy Evaluation Project (APEP).

Contributors: @“ai1scl”

First Contributor: https://github.com/âĂœai1sclâĂİ

Project Repository: https://github.com/SocialCatalystLab/auto-policy-evals