

Regression-Based Causal Inference for Person-Specific Effects in Within-Subject Randomization Designs

Chendong Li

rh6qm1@jmu.edu

James Madison University <https://orcid.org/0009-0007-3123-4706>

Wen Luo

Texas A&M University

Research Article

Keywords: Within-Subject Randomization Designs, Causal Inference, Individualized Treatment Effect

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11390517/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Regression-Based Causal Inference for Person-Specific Effects in Within-Subject Randomization Designs

Chendong Li^{1,2} and Wen Luo³

¹ Center for Assessment and Research Studies, James Madison University, Harrisonburg,
Virginia, USA

² Department of Graduate Psychology, James Madison University, Harrisonburg, Virginia, USA

³ Department of Educational Psychology, Texas A&M University, College Station, Texas, USA

Author Note

Correspondence concerning this article should be addressed to Chendong Li, Center for Assessment and Research Studies, James Madison University, Harrisonburg, Virginia, USA.

Email: rh6qml@jmu.edu.

Abstract

An intervention may benefit one person but have little effect on another, making person-specific treatment effects important in psychological and behavioral research. These effects are difficult to estimate in between-subject randomized experiments, where each person receives only one assigned condition, and the primary comparison yields an effect averaged across people. Within-subject randomization designs provide an alternative that repeatedly assign conditions to the same person, creating randomized within-person comparisons. Building on recent work on time-series experiments, we develop a regression-based causal framework for these designs. We define average immediate and lag-1 effects using potential outcomes and estimate them with an ordinary least square working regression whose treatment indicators are centered and scaled using known assignment probabilities. We extend the approach to multiple-arm interventions and restricted randomization. Two Monte Carlo studies found small bias for the proposed estimators across most conditions. In the multiple-arm study, inference improved with longer series. Under restricted randomization, probability adjustment reduced lag-1 bias, but short series yielded undercoverage and inflated Type I error with HAC standard errors; HC3 was more conservative and less powerful. Empirical demonstrations using a micro-randomized trial in health behavior and a restricted alternating treatment design in special education illustrate implementation and interpretation. The framework connects familiar regression tools with design-based causal inference for within-subject randomized studies.

Abstract

An intervention may benefit one person but have little effect on another, making person-specific treatment effects important in psychological and behavioral research. These effects are difficult to estimate in between-subject randomized experiments, where each person receives only one assigned condition, and the primary comparison yields an effect averaged across people. Within-subject randomization designs provide an alternative that repeatedly assign conditions to the same person, creating randomized within-person comparisons. Building on recent work on time-series experiments, we develop a regression-based causal framework for these designs. We define average immediate and lag-1 effects using potential outcomes and estimate them with an ordinary least square working regression whose treatment indicators are centered and scaled using known assignment probabilities. We extend the approach to multiple-arm interventions and restricted randomization. Two Monte Carlo studies found small bias for the proposed estimators across most conditions. In the multiple-arm study, inference improved with longer series. Under restricted randomization, probability adjustment reduced lag-1 bias, but short series yielded undercoverage and inflated Type I error with HAC standard errors; HC3 was more conservative and less powerful. Empirical demonstrations using a micro-randomized trial in health behavior and a restricted alternating treatment design in special education illustrate implementation and interpretation. The framework connects familiar regression tools with design-based causal inference for within-subject randomized studies.

Introduction

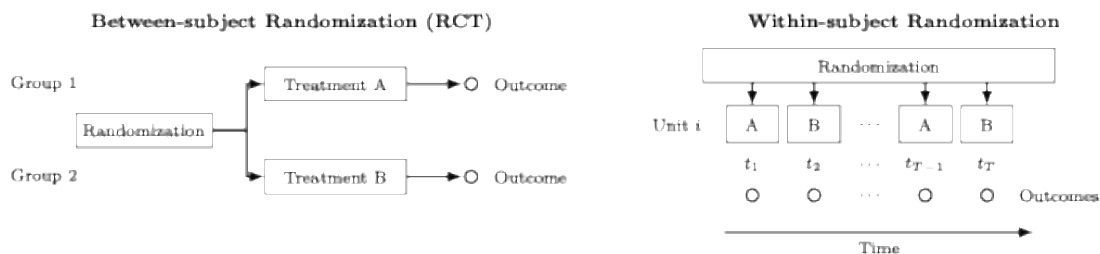
Making causal inferences in psychological and behavioral research is difficult because psychological phenomena and behavior are shaped by complex, interacting influences that may vary across individuals, contexts, and time (Bailey et al., 2024; Rohrer, 2018). Although randomized controlled trials (RCTs) are commonly treated as the tool to find causal evidence, they face challenges for answering psychological and behavioral questions. First, in practice, RCTs require recruiting large samples of participants to be assigned to separate treatment groups. However, recruitment feasibility is a persistent challenge, and prior reviews have found that many trials fail to reach their planned sample size or require extensions to do so (McDonald et al., 2006; Sully et al., 2013; Walters et al., 2017). Second, each participant in a between-subject RCT receives only one assigned condition, so the primary comparison estimates an effect averaged across participants rather than a treatment effect for a particular person. When the treatment effects vary, an intervention that is effective on average may have weaker, stronger, or even different direction for different individuals (Bryan et al., 2021; Bailey et al., 2024). This issue is greatly important as individualized and personalized approaches have received increasing attention in psychology and behavioral science, where researchers often seek to understand which treatment works for whom over what time progression (Molenaar, 2004; Barlow & Nock, 2009; Lillie et al., 2011; Davidson et al., 2014; Shaffer et al., 2018; Hekler et al., 2019). These considerations call for experimental designs that can evaluate treatment effects within an experimental unit while retaining the advantages of random assignment. Within-subject randomization designs provide one such approach.

Within-subject randomization designs refer to a class of experimental designs in which the same experimental unit is observed repeatedly, and treatment conditions are randomly

assigned to that unit over time. The unit of analysis may be a person, family, community, or any other entity that can serve as its own control (Epstein & Dallery, 2022; Heyvaert & Onghena, 2014). This design logic differs from that of a between-subject randomized experiment (e.g., a parallel-group RCT), in which different participants or units are randomly assigned to separate treatment groups to achieve comparability. In within-subject randomized designs, by contrast, the same experimental unit receives different treatment conditions at different times, with the timing of those conditions determined by random assignment, as illustrated in Figure 1. This class of designs is therefore especially relevant for psychological and behavioral research that seeks to evaluate individualized treatment effects while preserving advantages of random assignment (Kratochwill & Levin, 2010; Shadish et al., 2002).

Figure 1

Comparison of between-subject randomization (RCT) and within-subject randomization designs



It is important to note that the definition adopted in the present study is intentionally narrower than the broader category of traditional within-subjects designs (i.e., designs in which the same experimental unit are measured repeatedly across conditions or times (Kim, 2010)). Repeated measurement alone is not sufficient to isolate the effect of treatment (Shadish et al., 2002). The key feature of within-subject randomization designs is the random assignment of treatment conditions within the same unit across time. This randomization links the observed outcome sequence to a known assignment mechanism and provides the basis for internal validity

regarding treatment effectiveness (Edgington, 1980; Kratochwill & Levin, 2010). Thus, in the present study, the term within-subject randomization design refers not simply to repeated observation, but to repeated observation combined with randomized treatment assignment over time.

Relevant designs in psychological and behavioral sciences within the family of within-subject randomization designs are micro randomization trials (MRTs), N-of-1 trials with randomization, and randomized alternating treatment designs (ATDs) in single-case experimental design (SCED) (Klasnja et al., 2015; Lillie et al., 2011; Barlow & Hayes, 1979). These designs are not identical, but all involve repeated comparisons of randomly assigned treatment conditions within the same experimental unit. MRTs were developed primarily to inform the construction of just-in-time adaptive interventions (JITAI), which are digital interventions in which treatment components may be delivered repeatedly in daily life (Klasnja et al., 2015; Qian et al., 2022). In an MRT, each participant is repeatedly randomized to intervention options at times across the study period, so that investigators can evaluate whether an intervention affects the outcome for the individual and whether this effect varies across time or context. In the N-of-1 tradition, the goal is to determine the optimal intervention for that individual (Lillie et al., 2011). Randomized N-of-1 trials typically achieve the goal with internal validity through repeated crossover periods in which the order or sequence of treatment conditions is randomly assigned (Daza, 2018). In SCEDs, ATDs use a similar structure to address this: two or more conditions are rapidly alternated across closely spaced sessions or occasions to compare their relative effectiveness (Barlow & Hayes, 1979; Holcombe et al., 1994; Kazdin, 2021). Random assignment of treatment conditions over time is often used in ATDs to strengthen internal validity (Edgington, 1980, 1996; Kratochwill & Levin, 2010; Heyvaert &

Onghena, 2014). Thus, although MRTs, N-of-1 trials, and ATDs emerged from different methodological traditions and often serve different inferential goals, their randomized versions share a common design feature: treatment conditions vary within the same unit over time according to a known assignment mechanism.

Despite the promise of within-subject randomization designs, current analytic approaches focus on summarizing observed differences between treatment conditions. Researchers often rely on visual analysis or simple statistical analysis, such as mean differences, nonoverlap indices, t tests, or simple regression-based summaries (Manolov & Onghena, 2018; Shaffer et al., 2018; De Carvalho et al., 2026). These approaches can be useful for description and exploratory analysis in applied settings. However, they may not clearly separate differences attributable to the current treatment exposure from differences associated with time trends, serial dependence, or prior treatment exposure (Barlow & Hayes, 1979; Shamseer et al., 2015; Kazdin, 2021; De Carvalho et al., 2026; Luo & Li, 2026).

To more rigorously investigate causal effects in within-subject randomized designs, recent methodological work uses the counterfactual framework (Rubin, 1974; Holland, 1986) to formally define causal estimands (i.e., the causal effects of interest) and proposed estimators using the technique of inverse probability weighting (IPW) (Daza, 2018; Bojinov & Shephard, 2019). Luo and Li (2026) investigated the performance of IPW-based estimators for the average immediate treatment effect and the average lag-1 effect in ATDs. The findings indicate that IPW-based estimators outperformed the traditional methods such as mean difference, percent of nonoverlap data (PND-W), and piecewise regression, when lagged effects are present, randomization is restricted, and series length is moderate. Their work is especially relevant to psychological and behavioral research because it evaluates these estimators under short to

moderate series lengths (i.e., 12 to 24 time points) and designs with restricted randomizations. Restricted randomization are often used in ATDs when the number of measurement occasions is limited, to avoid long runs of the same condition, increase practical feasibility, or reduce concerns about poorly balanced treatment sequences (What Works Clearinghouse [WWC], 2022).

More recently, Lin and Ding (2025) offer a regression-based estimator for average immediate and lagged effects in the within-subject randomized design. Instead of applying inverse probability weights to observations, their method transforms the treatment indicators, allowing treatment effects to be estimated using ordinary least squares (OLS) regression while retaining a design-based causal interpretation. Relative to the IPW estimator, this approach is more appealing to applied researchers because it keeps the causal inference in a familiar regression framework, avoiding the need to construct weights and compute contrast-coded and weighted outcomes. The regression-based estimator may also be more efficient than the design-based IPW estimator, because it estimates the average immediate and lagged effects jointly through multiple predictors in a single regression model, which can reduce the residual variance, and hence the variance of the regression coefficients. Despite these advantages, the original regression-based estimator is developed for a relatively narrow setting with only two treatment conditions and randomization without restrictions. As a result, the method cannot be directly applied to designs with more than two treatment conditions and/or restricted randomization schemes. Broadening the framework to cover these cases is necessary for making regression-based causal inference both methodologically rigorous and practically useful in within-subject randomization studies.

Present Study

The present study addresses the methodological gap by extending the regression-based framework and evaluating its performance through simulation and empirical demonstration. The study has four aims. First, following Lin and Ding (2025), we define average immediate and lag-1 causal effects for within-subject randomized designs using potential outcome framework and show how OLS recover the causal effects under a known treatment assignment scheme. Second, we extend the regression-based formulation to design features common in psychological and behavioral applications: (1) intervention with more than two conditions and (2) randomization with restrictions. Third, we evaluate the performance of the estimators through simulation studies. Finally, we demonstrate their implementation using empirical examples, showing how the resulting coefficients can be interpreted as causal effect estimates in real world applications.

The remainder of the paper is organized as follows. We first introduce the regression-based causal inference framework for within-subject randomization designs. We then extend the framework to (1) more than two treatment conditions and (2) restricted randomization. Next, we conduct simulation studies to evaluate the performance of the proposed estimator under different conditions. We then provide two empirical demonstrations using data from an MRT in health behavior and a randomized ATD in special education. Finally, we discuss the findings, methodological implications, limitations, and directions for future research.

Regression-based Causal Estimation under With-subject Randomization

The Causal Estimands

Consider a single experimental unit observed over T occasions. Let W_t denote the treatment condition indicator at occasion t , and let $W_{1:t} = (W_1, \dots, W_t)$ denote the treatment series up to occasion t . Let $Y_t(W_{1:t})$ denote the potential outcome at occasion t under treatment

series $W_{1:t}$. This notation indicates that in a within-subject randomization design, the outcome at a given occasion may depend not only on the treatment assigned at that occasion W_t but also on prior treatment exposure $W_{1:t-1}$.

For a two-condition within-subject randomization design, let $W_t = 1$ denote the treatment condition and $W_t = 0$ denote the comparison condition. The time-specific immediate effect at occasion t is given by

$$\delta_t^{(0)} = E[Y_t(W_{1:t-1}, 1) - Y_t(W_{1:t-1}, 0)], \quad (1)$$

where the expectation is taken over the prior treatment history generated by the randomization design. This estimand is focused on the question: at occasion t , what would be the effect of assigning treatment rather than comparison, averaging over the possible histories that could have occurred before t .

The time-specific effect in Equation (1) defines the causal contrast at a particular occasion. However, researchers are usually interested in the treatment effect across a period of repeated exposure rather than at single isolated occasion which may be sensitive to session-specific events. Therefore, we average these time-specific contrasts over a prespecified analysis window. This window defines the series over which the treatment effect is intended to be summarized. The average immediate effect over the analysis window $\mathcal{T}_0 = \{k, \dots, l\}$ is

$$\bar{\Delta}^{(0)} = \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \delta_t^{(0)}. \quad (2)$$

The time-specific lag-1 effect at occasion t compares treatment with comparison at occasion $t - 1$, holding the treatment at occasion t the same within each comparison. These

comparisons are then averaged over earlier treatment histories and the possible treatment assignments at occasion t , according to the randomization design. The lag-1 effect is defined for $t \geq 2$.

$$\delta_t^{(1)} = E \left[\sum_{W_t \in \{0,1\}} P(W_t = w | W_{1:t-2}) \times \{Y_t(W_{1:t-2}, 1, w) - Y_t(W_{1:t-2}, 0, w)\} \right]. \quad (3)$$

In Equation (3), $P(W_t = w | W_{1:t-2})$ is the design probability of treatment assignment at occasion t used to average over the possible current treatment assignments. For each value of w , the two potential outcomes differ only in the treatment assigned at occasion $t - 1$ and both have the same current treatment w . The contrast are then averaged over w and over treatment histories through occasion $t - 2$. Thus, the lag-1 effect captures the effect of the treatment at the previous occasion on the current outcome. As with the time-specific immediate effect, we summarize the lag-1 effects over a prespecified analysis window. The average lag-1 effect over the analysis window $\mathcal{T}_1 = \{t: k \leq t \leq l, t \geq 2\}$ is given similarly with average immediate effect as shown below

$$\bar{\Delta}^{(1)} = \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} \delta_t^{(1)}. \quad (4)$$

Building on the same idea, a lag- k effect compares the outcome at occasion t under two treatment histories that differ only in the treatment assigned at occasion $t - k$. All other treatment assignments are held the same within each comparison. These contrasts can then be averaged over earlier histories and over a common distribution of the assignments between occasions $t - k + 1$ to t . In the present study, we focus on the immediate and lag-1 effects as

they correspond to the most common causal questions in within-subject randomization designs (Luo & Li, 2026).

Regression-based Estimator

After defining the causal estimands, the next step is to estimate them from the observed outcome sequence. The main idea is to use the known randomization scheme to construct treatment predictors that represent random deviations from the expected treatment assignment probability. Let $p_t = P(W_t = 1 \mid W_{1:t-1})$ denote the known conditional probability of receiving treatment at occasion t , given the prior treatment history. Under an unrestricted randomization for within-subject studies, treatment assignment at each time point is independent of history, so $p_t = P(W_t = 1)$. The variance of W_t is $v_t = p_t(1 - p_t)$. When $0 < p_t < 1$, the transformed treatment indicator at occasion t is

$$\tilde{W}_t = \frac{W_t - p_t}{v_t} \quad (5)$$

Similar, let $p_{t-1} = P(W_{t-1} = 1 \mid W_{1:t-2})$ be conditional probability of receiving treatment at occasion $t - 1$, given the prior treatment history. The variance of W_{t-1} is $v_{t-1} = p_{t-1}(1 - p_{t-1})$. When $0 < p_{t-1} < 1$, the transformed treatment indicator at occasion $t - 1$ is

$$\tilde{W}_{t-1} = \frac{W_{t-1} - p_{t-1}}{v_{t-1}} \quad (6)$$

For both $\tilde{W}_t$ and $\tilde{W}_{t-1}$, the numerator centers the observed assignment around its design expectation, and the denominator scales these deviations by the corresponding assignment variance, v_t or v_{t-1} . Then $\tilde{W}_t$ and $\tilde{W}_{t-1}$ are used as predictors in the working OLS regression to estimate the immediate and lag-1 effects respectively:

$$Y_t = b_0 \tilde{W}_t + b_1 \tilde{W}_{t-1} + e_t; t > 1 \quad (7)$$

The coefficient $\hat{b}_0$ summarizes the relationship between the randomized deviation in the treatment assignment at occasion t and the outcome at occasion t (i.e., proportional to the average immediate effect), whereas $\hat{b}_1$ summarizes the relationship between the randomized deviation in the treatment assignment at occasion $t - 1$ and the outcome at occasion t (i.e., proportional to the average lag-1 effect). Because the treatment predictors are scaled by the assignment variance, these coefficients must be rescaled to represent the true treatment effects using the following weights defined

$$c_0 = \left[\frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \frac{1}{v_t} \right] \quad (8)$$

$$c_1 = \left[\frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} \frac{1}{v_{t-1}} \right] \quad (9)$$

The regression-based causal estimators are then

$$\hat{\Delta}^{(0)} = c_0 \hat{b}_0 \quad (10)$$

$$\hat{\Delta}^{(1)} = c_1 \hat{b}_1 \quad (11)$$

When $p_t = 0.5$ for all t , $v_t = 0.25$, the transformation is especially simple $\hat{\Delta}^{(0)} = 4\hat{b}_0$ and

$\hat{\Delta}^{(1)} = 4\hat{b}_1$. It should be noted that the regression model does not imply that the outcome is generated by a linear model. It is not intended to explain the data generation mechanism. It is a working model that uses the randomized treatment variation to estimate the causal estimands defined above.

Assumptions for Causal Interpretation

Lin and Ding (2025) establish consistency and asymptotic normality for the OLS-based estimator under their design and regularity conditions. We briefly summarize these assumptions here. The causal interpretation of the regression-based estimator relies on four assumptions:

1. Sequential consistency. The treatment conditions must be well defined, and the observed outcome must correspond to the potential outcome under the treatment history actually received. That is, if the realized treatment history through occasion t is $w_{1:t}$, then $Y_t^{obs} = Y_t(w_{1:t})$.
2. Sequential positivity. To estimate an immediate effect at occasion t , both treatment and comparison must be possible at that occasion, meaning $0 < p_t < 1$. Similarly, to estimate a lag-1 effect at occasion t , the previous assignment W_{t-1} must have been randomized, so $0 < p_{t-1} < 1$. If a treatment condition is assigned deterministically such that $p_t = 0$ or 1 , the corresponding variance v_t is zero and the transformed indicator $\tilde{W}_t$ is undefined. In an unrestricted randomization scheme, positivity is guaranteed by design. In contrast, under a restricted randomization scheme, positivity may fail to hold for on some occasions.
3. Sequential exchangeability. On each occasion, conditional on the observed history available before treatment assignment, the assigned treatment is independent of the current and future potential outcomes under each treatment history. In the randomized designs considered here, this condition is ensured by implementing the prespecified assignment scheme.
4. No anticipation. The potential outcome at occasion t may depend on the treatment history up to and including occasion t , but it should not depend on future treatment assignments.

Extensions of Regression-based Causal Estimation

The original regression-based causal estimator was developed for binary treatment conditions under unrestricted randomization. In this section, we extend the framework to two design features common in psychological and behavioral research: intervention designs with more than two treatment conditions and restricted randomization.

Extension I: Multiple-arm intervention designs

Many within-subject randomization designs in psychological and behavioral research involve more than two treatment conditions. For example, a study may compare several active intervention components with a common comparison condition or compare several treatment variants against a no-treatment or usual-practice condition (e.g., Klasnja et al., 2019). In this setting, the scalar treatment indicator W_t used in the previous section must be extended to a treatment-indicator vector.

For a J -condition within-subject randomization design, let $W_t = 0$ denote the comparison condition and $W_t = j$ denote treatment condition j , where $j = 1, \dots, J - 1$. As before, let $W_{1:t} = (W_1, \dots, W_t)$ denote the treatment series up to occasion t , and let $Y_t(W_{1:t})$ denote the potential outcome at occasion t under treatment series $W_{1:t}$. Treatment assignment at each occasion is independent of history under the unrestricted randomization considered in this extension. Let $p_{tj} = P(W_t = j \mid W_{1:t-1}) = P(W_t = j)$ denote the known design probability for condition j .

The comparison-condition probability is $p_{t0} = 1 - \sum_{j=1}^{J-1} p_{tj}$. We assume that all J conditions have positive assignment probabilities.

Define the treatment-indicator vector $D_t = (I(W_t = 1), \dots, I(W_t = J - 1))^T$ and the probability vector $p_t = (p_{t1}, \dots, p_{t,J-1})^T$. Here, $I(\cdot)$ is an indicator function. The entries of D_t indicate which treatment condition was assigned at occasion t , with $W_t = 0$ serving as the omitted comparison condition. In this extension, D_t and p_t are $(J - 1) \times 1$ vectors.

The time-specific immediate effect of treatment condition j relative to the comparison condition at occasion t is given by

$$\delta_{t,j-0}^{(0)} = E[Y_t(W_{1:t-1}, j) - Y_t(W_{1:t-1}, 0)]. \quad (12)$$

As in Equation (1), the expectation is taken over the prior treatment history generated by the randomization design. Averaging these time-specific effects over the prespecified analysis window $\mathcal{T}_0$ gives the average immediate effect of treatment condition j relative to the comparison condition:

$$\Delta_{j-0}^{(0)} = \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \delta_{t,j-0}^{(0)}. \quad (13)$$

Similarly, the time-specific lag-1 effect at occasion t compares treatment condition j with the comparison condition at the previous occasion while holding the current treatment fixed. The lag-1 effect is defined for $t \geq 2$:

$$\begin{aligned} \delta_{t,j-0}^{(1)} = E & \left[\sum_{w=0}^{J-1} P(W_t = w \mid W_{1:t-1}) \right. \\ & \times \{Y_t(W_{1:t-2}, j, w) - Y_t(W_{1:t-2}, 0, w)\}]. \end{aligned} \quad (14)$$

This definition compares two histories that differ only at $t - 1$, while the current treatment is held the same inside each contrast. The design probabilities are used to average over

the possible current treatment assignments. Under unrestricted randomization, $P(W_t = w \mid W_{1:t-1}) = p_{tw}$, so the same weight is used for both potential outcomes in each contrast. As with the time-specific immediate effect, we summarize the lag-1 effects over the prespecified analysis window $\mathcal{T}_1$:

$$\Delta_{j-0}^{(1)} = \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} \delta_{t,j-0}^{(1)}. \quad (15)$$

For convenience, we organize the target causal estimands in a vector $\Delta = (\Delta_{1-0}^{(0)}, \dots, \Delta_{(J-1)-0}^{(0)}, \Delta_{1-0}^{(1)}, \dots, \Delta_{(J-1)-0}^{(1)})^\top$. Thus, there are $2(J-1)$ target effects: $J-1$ average immediate effects and $J-1$ average lag-1 effects, each relative to the same comparison condition. In the subscripts, $j-0$ denotes a contrast between condition j and condition 0.

To estimate these effects using OLS, we first center and scale the treatment-indicator vectors D_t and D_{t-1} using the known assignment covariance matrices:

$$\tilde{D}_t = \Sigma_t^{-1}(D_t - p_t). \quad (16)$$

$$\tilde{D}_{t-1} = \Sigma_{t-1}^{-1}(D_{t-1} - p_{t-1}). \quad (17)$$

Analogous to the binary transformations in Equations (5) and (6), subtracting p_t centers the observed assignment around its design expectation. The covariance matrix Σ_t extends the assignment variance v_t to multiple treatment indicators:

$$\Sigma_t = \text{Var}(D_t) = \text{diag}(p_t) - p_t p_t^\top. \quad (18)$$

The diagonal entries of Σ_t are $p_{tj}(1 - p_{tj})$, and the off-diagonal entries are $-p_{tj}p_{tk}$ for $j \neq k$. Because only one condition can be assigned at each occasion, the treatment indicators are

negatively correlated. Multiplication by Σ_t^{-1} accounts for both their variances and these covariances. The matrix is invertible when all treatment conditions and the comparison condition have positive probabilities. The transformed $\tilde{D}_t$ is a $(J - 1) \times 1$ vector representing randomized deviations from the expected treatment assignments.

After the transformation, the working OLS regression is specified as

$$Y_t = \tilde{D}_t^\top b_0 + \tilde{D}_{t-1}^\top b_1 + e_t, t > 1. \quad (19)$$

Here, $b_0 = (b_{0,1}, \dots, b_{0,J-1})^\top$ and $b_1 = (b_{1,1}, \dots, b_{1,J-1})^\top$ are $(J - 1) \times 1$ coefficient vectors for the transformed current and lag-1 treatment indicators, respectively. As in the binary case, these coefficients must be rescaled to represent the treatment effects. The matrices used to transform the regression coefficients back to the causal-effect scale are defined as

$$M_0 = \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \Sigma_t^{-1}. \quad (20)$$

$$M_1 = \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} \Sigma_{t-1}^{-1}. \quad (21)$$

Let $\hat{b} = (\hat{b}_0^\top, \hat{b}_1^\top)^\top$. The regression-based causal estimator is then

$$\hat{\Delta} = \begin{pmatrix} M_0 & 0 \\ 0 & M_1 \end{pmatrix} \hat{b}. \quad (22)$$

The first $J - 1$ elements of $\hat{\Delta}$ estimate the average immediate effects, and the remaining $J - 1$ elements estimate the average lag-1 effects. When the assignment probabilities are constant across occasions, $M_0 = M_1 = \Sigma^{-1}$. When $J = 2$, the vector and matrix formulation reduces to the binary transformation and rescaling introduced above. When $J = 3$, it gives two immediate

effects and two lag-1 effects. Simulation Study I evaluates this three-condition application of the multiple-arm extension.

The multiple-arm extension relies on the same four assumptions stated above: sequential consistency, sequential positivity, sequential exchangeability, and no anticipation. Here, positivity requires every condition to have positive assignment probability at the relevant current and preceding occasions, making Σ_t invertible. The implemented randomization supplies exchangeability. As above, OLS is a working regression.

Extension II: Restricted Randomization

Another design feature in within-subject randomization studies is to use restrictions on the treatment sequence to avoid long runs of the same condition. This feature is common in randomized ATDs in SCEDs because researchers often want treatment conditions to alternate frequently enough to yield sufficient crossovers to demonstrate treatment effects (WWC, 2022). For example, Eilers and Hayes (2015) used a two-condition ATD in which no more than two sessions of the same condition were administered consecutively.

To aid comprehension and reduce notational burden, we present the extension for a two-condition design with a restricted randomization scheme that allows no more than two consecutive sessions of the same condition. As before, let $W_t = 1$ denote the treatment condition and $W_t = 0$ denote the comparison condition. Let $p_t = P(W_t = 1 \mid W_{1:t-1})$ denote the known conditional probability of receiving treatment at occasion t , given the prior treatment history. Let q denote the probability of receiving treatment when the restriction does not determine the

assignment, where $0 < q < 1$. The first two assignments have probability $p_1 = p_2 = q$. For $t \geq 3$, the conditional probability is given by

$$p_t = \begin{cases} 1, & \text{if } W_{t-2} = 0 \text{ and } W_{t-1} = 0, \\ 0, & \text{if } W_{t-2} = 1 \text{ and } W_{t-1} = 1, \\ q, & \text{otherwise.} \end{cases} \quad (23)$$

This restriction causes the treatment assignment at some occasions to be deterministic: when $p_t = 0$ or $p_t = 1$, the treatment at occasion t is fixed by the restriction rule and does not provide randomized information for the immediate treatment contrast. The assignment variance is still given by $v_t = p_t(1 - p_t)$. The transformed current treatment indicator is then defined as

$$\tilde{W}_t = \begin{cases} \frac{W_t - p_t}{v_t}, & \text{if } v_t > 0, \\ 0, & \text{if } v_t = 0. \end{cases} \quad (24)$$

The same definition is used for the transformed treatment indicator at occasion $t - 1$, replacing p_t and v_t with p_{t-1} and v_{t-1} . The zero coding indicates that a deterministic assignment provides no randomized deviation from its conditional expectation; it does not restore positivity for that occasion.

The working OLS regression under restricted randomization also differs from the model under unrestricted randomization. Because the assignment probabilities depend on prior treatment history, we propose a working regression that includes both p_t and p_{t-1} :

$$Y_t = \alpha + b_0 \tilde{W}_t + b_1 \tilde{W}_{t-1} + \gamma_0 p_t + \gamma_1 p_{t-1} + e_t. \quad (25)$$

Here, p_t and p_{t-1} are design-based adjustment terms that account for the assignment dependence created by the restriction rule. Their inclusion follows from the identities $W_t = p_t +$

$q(1 - q)\tilde{W}_t$ and $W_{t-1} = p_{t-1} + q(1 - q)\tilde{W}_{t-1}$, which also hold on deterministic occasions.

Thus, a constant additive immediate effect contributes a term involving p_t , and a constant additive lag-1 effect contributes a term involving p_{t-1} . The transformed variables $\tilde{W}_t$ and $\tilde{W}_{t-1}$ continue to represent randomized deviations from the conditional assignment probabilities.

When $q = 0.5$, treatment and comparison are equally likely whenever the assignment is randomized. A forced assignment at occasion $t - 1$ is followed by a randomized assignment at occasion t . Thus, whenever p_{t-1} differs from 0.5, $p_t = 0.5$. The transformed treatment indicators are centered on their assignment probabilities and have expected value zero over the assignments allowed by the design. Together, these properties make the varying part of p_{t-1} uncorrelated, in design expectation, with p_t and both transformed treatment indicators, and the intercept absorbs its constant part. Thus, omitting p_{t-1} does not change the population treatment coefficients. We retain p_t because it depends on the previous treatment assignment and can be related to the lag-1 predictor. This gives the reduced working regression with only the current probability adjustment:

$$Y_t = \alpha + b_0\tilde{W}_t + b_1\tilde{W}_{t-1} + \gamma p_t + e_t. \quad (26)$$

This is the reduced formulation evaluated in Simulation Study II. The simplification concerns population coefficients, and it does not imply that the two regressions produce identical coefficients in each sample or that the reduced estimator is exactly unbiased in finite samples.

When $q \neq 0.5$, both probability terms should be retained.

Within the prespecified analysis window, let the eligible occasions for rescaling be $\mathcal{T}_0 = \{t: v_t > 0\}$ and $\mathcal{T}_1 = \{t: t \geq 2, v_{t-1} > 0\}$. The working regression can retain occasions with

deterministic assignments using the zero coding above. The coefficient b_0 summarizes the relationship between the randomized deviation in treatment assignment at occasion t and the outcome at that occasion, whereas b_1 summarizes the corresponding relationship for treatment assignment at occasion $t - 1$. Because the transformed treatment indicators are divided by their positive conditional assignment variances, these coefficients must be rescaled to represent treatment effects. As in Equations (8) and (9), define

$$c_0 = \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \frac{1}{v_t}. \quad (27)$$

$$c_1 = \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} \frac{1}{v_{t-1}}. \quad (28)$$

The regression-based causal estimators are then $\hat{\Delta}^{(0)} = c_0 \hat{b}_0$ and $\hat{\Delta}^{(1)} = c_1 \hat{b}_1$. When $q = 0.5$, every positive assignment variance equals 0.25. Hence, $c_0 = c_1 = 4$, and the estimated immediate and lag-1 effects are $4\hat{b}_0$ and $4\hat{b}_1$, respectively. The factor of four converts the coefficients from the transformed-predictor scale to the causal-effect scale, and deterministic occasions are retained in the working regression but are excluded from the rescaling sums.

The assumptions of sequential consistency, sequential exchangeability, and no anticipation remain as stated above. However, the restriction makes some assignments deterministic, so sequential positivity does not hold for every treatment history. The zero-coded transformed indicators retain those occasions in the working regression but do not supply an immediate or lag-1 contrast where assignment was forced. The constant additive effects assumption allow the effects on eligible histories to extend to those unsupported contrasts. Without that additional assumption, the causal interpretation is limited to histories at which the

relevant assignment remains randomized. The derivation and its conditions are provided in the supplemental materials.

Inference and Standard Error Estimation

After fitting the working OLS model, uncertainty was estimated using robust covariance estimators. We considered two standard error options: Newey–West heteroskedasticity- and autocorrelation-consistent (HAC) standard errors and HC3 heteroskedasticity-robust standard errors.

The Newey–West HAC standard error estimator (Newey & West, 1987) is useful in time-series settings because the estimating terms from the working regression may remain serially dependent. Although the regression includes transformed current and lag-1 treatment indicators, it does not require the remaining variation to be independent over time. Serial dependence may still arise from prior outcomes, longer-lag carryover beyond lag 1, or other unmodeled temporal factors. Following common practice, the HAC bandwidth was set to $\lfloor T^{1/4} \rfloor$, where T denotes the number of occasions included in the fitted working regression (Greene, 2003; Lin & Ding, 2025).

The second option was the HC3 robust standard error estimator (Long & Ervin, 2000). HC3 does not explicitly adjust for serial dependence, but it is often useful in small-sample regression settings because it reduces downward bias in heteroskedasticity-robust variance estimation when leverage is non-negligible. This feature is relevant for psychological within-subject randomization designs, which often have relatively short series lengths.

For the binary and restricted-randomization estimators, standard errors for the causal-effect estimates were obtained by applying the same rescaling used for the point estimates. That is, after the OLS coefficients were transformed back to the causal-effect scale, their standard

errors were transformed using the corresponding scaling factors. For the multiple-arm estimator, the covariance matrix of the regression coefficients was transformed using the same linear transformation applied to the coefficient vector. This ensured that both the point estimates and their standard errors were reported on the causal-effect scale. Confidence intervals and p-values were then computed from the transformed estimates and transformed standard errors.

Statistical Simulation

To evaluate the proposed regression-based causal estimator under conditions relevant to psychological within-subject randomization designs, we conducted two Monte Carlo simulation studies. The first simulation evaluated Extension I in a three-condition intervention design, providing an application of the general multiple-arm formulation. The second simulation evaluated the proposed estimator under restricted randomization. We did not include a separate unrestricted two-condition simulation because that setting corresponds to the baseline formulation already introduced in the preceding theoretical section. The three-condition simulation evaluates the multiple-arm formulation with two treatment conditions and a common comparison condition.

Simulation Study I

Data Generation and Conditions

The first simulation study evaluated Extension I in a three-condition within-subject randomization design, corresponding to $J = 3$. The comparison condition was coded as $W_t = 0$ and labeled condition A. The two treatment conditions were coded as $W_t = 1$ and $W_t = 2$ and labeled conditions B and C, respectively. On each occasion, one of the three conditions was

independently assigned with known design probabilities $P(W_t = 0) = .40$, $P(W_t = 1) = .30$, and $P(W_t = 2) = .30$. The observed outcome was generated from the following model:

$$Y_t = \beta_1 \log(t) + \delta_{B0} I(W_t = 1) + \delta_{C0} I(W_t = 2) + \delta_{B1} I(W_{t-1} = 1) + \delta_{C1} I(W_{t-1} = 2) + \varepsilon_t. \quad (29)$$

where $\varepsilon_t \sim N(0,1)$, and condition A was the comparison condition. The parameters δ_{B0} and δ_{C0} represent the true immediate treatment effects of conditions B and C, respectively, compared with condition A. The parameters δ_{B1} and δ_{C1} represent the corresponding lag-1 effects. Under the constant additive effects used here, these four parameters equal $\Delta_{1-0}^{(0)}$, $\Delta_{2-0}^{(0)}$, $\Delta_{1-0}^{(1)}$, and $\Delta_{2-0}^{(1)}$, respectively. The log-time trend was included to evaluate performance when the outcome changes over time for reasons unrelated to treatment. For inference, two robust standard error options were examined: Newey–West HAC standard errors and HC3 robust standard errors. HAC standard errors account for possible heteroskedasticity and serial dependence in the working-regression estimating terms, whereas HC3 standard errors provide a finite-sample heteroskedasticity-robust option when leverage may be non-negligible.

The simulation conditions crossed the following factors: time trend, immediate effect of condition B, immediate effect of condition C, lag-1 effect of condition B, lag-1 effect of condition C, and series length T . These conditions produced 160 simulation conditions. For each condition, 1,000 datasets were generated. The detailed simulation conditions for the Simulation I is shown in Table 1.

Table 1*Simulation Conditions for the Three-Condition Design*

Factor	Values
Time trend (β_1)	(0, 1)
Immediate effect of Condition B (δ_{B0})	(0, 3)
Immediate effect of Condition C (δ_{C0})	(0, 2)
Lag-1 effect of Condition B (δ_{B1})	(0, 1)
Lag-1 effect of Condition C, (δ_{C1})	(0, 1)
Series length (T)	(20, 50, 100, 150)

Analytical Methods

For each simulated dataset, the multiple-arm estimator was implemented for three conditions using the procedure described in Extension I. The treatment indicators for conditions B and C were transformed using the assignment covariance matrix in Equations (16)–(18), and the working OLS model in Equation (19) was fitted to each simulated dataset:

$$Y_t = \tilde{D}_t^\top b_0 + \tilde{D}_{t-1}^\top b_1 + e_t. \quad (30)$$

The resulting coefficient vector was then mapped back to the causal-effect scale using the block-matrix rescaling in Equation (22). With conditions A, B, and C corresponding to codes 0, 1, and 2, the four estimates were denoted by $\hat{\Delta}_{1-0}^{(0)}$, $\hat{\Delta}_{2-0}^{(0)}$, $\hat{\Delta}_{1-0}^{(1)}$, and $\hat{\Delta}_{2-0}^{(1)}$. These estimate the immediate and lag-1 effects of condition B versus condition A and condition C versus condition A.

For inference, both Newey–West HAC and HC3 robust standard errors were applied to the fitted working regression. The coefficient covariance matrix was transformed using the same linear transformation used for the point estimates.

Simulation Study II

Data Generation and Conditions

The second simulation study evaluated the proposed estimator under restricted randomization. The design included a treatment condition and a comparison condition, $W_t \in \{0,1\}$, but the treatment sequence was restricted so that no condition could be assigned more than two times consecutively. This restriction reflects a common design choice in short randomized alternating-treatment designs, where researchers may wish to avoid long runs of the same condition. The purpose was to examine whether the proposed probability-adjusted regression-based estimator remains unbiased when the treatment assignment mechanism depends on prior assignments.

The first two occasions were randomized with $p_1 = p_2 = .50$. For $t \geq 3$, the conditional probability depended on the two previous assignments:

$$p_t = P(W_t = 1 \mid W_{1:t-1}) = \begin{cases} 1, & \text{if } W_{t-2} = 0 \text{ and } W_{t-1} = 0, \\ 0, & \text{if } W_{t-2} = 1 \text{ and } W_{t-1} = 1, \\ .50, & \text{otherwise.} \end{cases} \quad (31)$$

The outcome was generated using the same basic structure as in Simulation Study I:

$$Y_t = \beta_1 \log(t) + \delta_0 W_t + \delta_1 W_{t-1} + \varepsilon_t, \quad (32)$$

where $W_0 = 0$ and $\varepsilon_t \sim N(0,1)$.

The simulation conditions crossed time trend, immediate effect, lag-1 effect, and series length T . These conditions produced 48 simulation cells. For each cell, 1,000 datasets were generated. The detailed simulation conditions for the Simulation II are shown in Table 2.

Table 2

Simulation Conditions for the Restricted-randomization Designs

Factor	Values
Time trend (β_1)	(0, 1)
Immediate effect (δ_0)	(0, 2, 3, 4)
Lag-1 effect (δ_1)	(0, 1)
Series length (T)	(15, 30, 60)

Analytical Methods

For each simulated dataset, the conditional assignment probabilities were computed from the restricted randomization rule. The transformed treatment indicators were then constructed using the restricted-randomization transformation defined above, with deterministic occasions coded as zero so that they did not contribute randomized treatment variation.

Two working regression estimators were compared. The first was a non-probability-adjusted working regression,

$$Y_t = b_0 \tilde{W}_t + b_1 \tilde{W}_{t-1} + e_t, \quad (33)$$

which used the transformed current and lag-1 treatment indicators but did not include the additional design-based adjustment term. The second was the reduced probability-adjusted working regression for equal assignment probabilities on randomized occasions, as shown in Equation (26):

$$Y_t = \alpha + b_0 \tilde{W}_t + b_1 \tilde{W}_{t-1} + \gamma p_t + e_t. \quad (34)$$

This comparison was included to evaluate whether adding p_t is necessary when the randomization probability depends on prior treatment history. For both estimators, the immediate and lag-1 coefficients were rescaled to the causal-effect scale using the same logic as Equations (8)–(11), but only over occasions satisfying the relevant positivity condition. HAC and HC3 robust standard errors were applied and transformed using the same scaling factors as the point estimates.

Evaluation Metrics

The same evaluation criteria were used across the two simulation studies. For each simulation condition, point-estimation accuracy was assessed using empirical bias, relative bias, and root mean squared error (RMSE). Empirical bias was defined as $\text{Bias} = \bar{\hat{\Delta}} - \Delta$, where $\bar{\hat{\Delta}}$ is the average estimated effect across replications and Δ is the true effect. Relative bias was computed by dividing empirical bias by the true effect and was reported only when the true effect was nonzero. RMSE summarized the overall estimation error by combining both bias and variability across replications. Inferential performance was evaluated using confidence interval coverage, empirical Type I error, and empirical power. For each replication, a 95% Wald-type confidence interval was constructed using the estimated standard error from the fitted model.

Coverage was defined as the proportion of replications in which the confidence interval contained the true effect. Empirical Type I error was computed as the rejection rate at the .05 level when the true effect was zero, whereas empirical power was computed as the rejection rate when the true effect was nonzero. For each simulation study, these metrics were reported separately for the immediate effect and the lag-1 effect. In the three-condition study, the metrics were reported separately for each active condition relative to the reference condition.

Transparency and Openness

The R code for the two simulation studies and empirical demonstrations is available at <https://github.com/cliatca/within-subject-causal-inference>. The values used for the Jacob demonstration are included in the demonstration script. The HeartSteps data file needed to rerun that demonstration is not currently included in the repository.

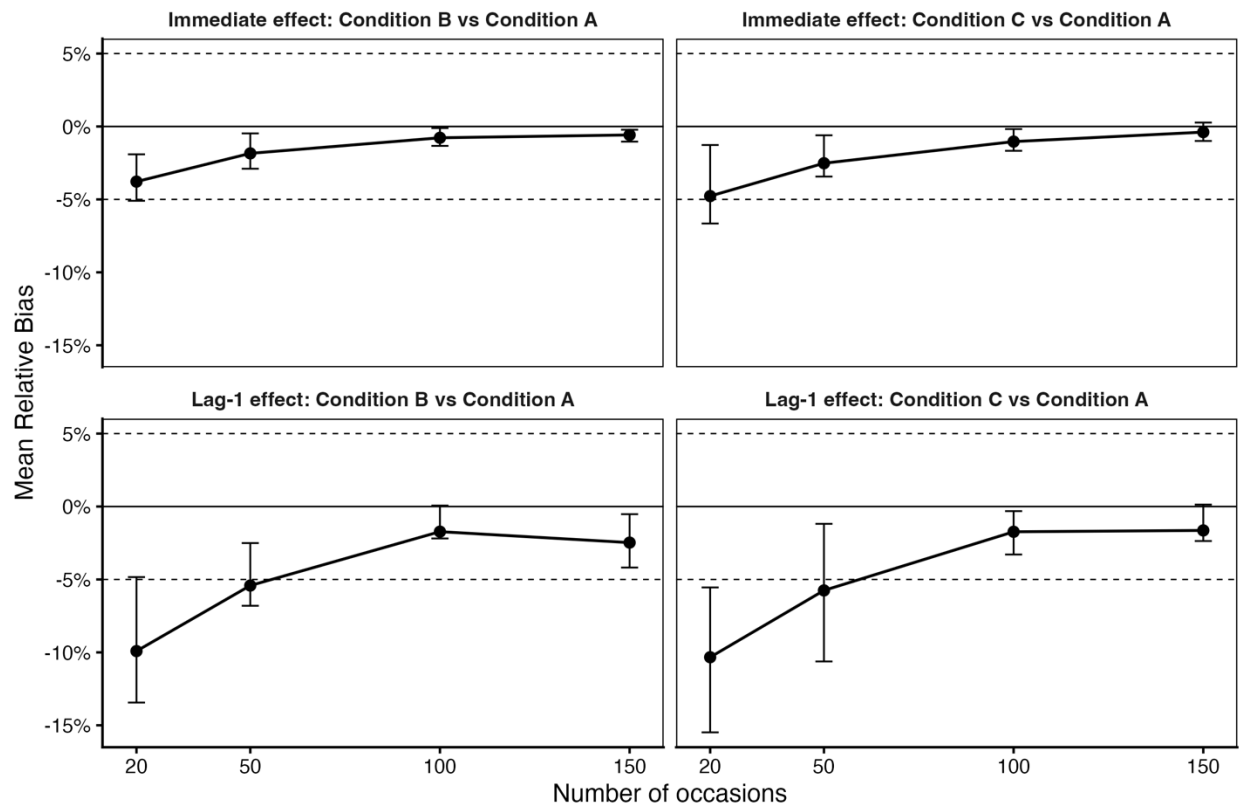
Simulation Results

Simulation Study I

The relative bias results showed that the three-condition transformed-regressor estimator recovered the causal effects with small bias across most conditions. Relative bias was largest when the number of occasions was smallest, especially for the lag-1 effects, but it decreased as the series length increased. For the immediate effects, relative bias was generally close to zero across series lengths and moved further toward zero as T increased. For the lag-1 effects, relative bias was more variable in the shorter series but also decreased with larger T . Overall, these results suggest that the estimator was approximately unbiased for both immediate and lag-1 effects, with greater finite-sample variability when the series length was short. Figure 2 shows the mean relative bias for each causal contrast across series lengths.

Figure 2

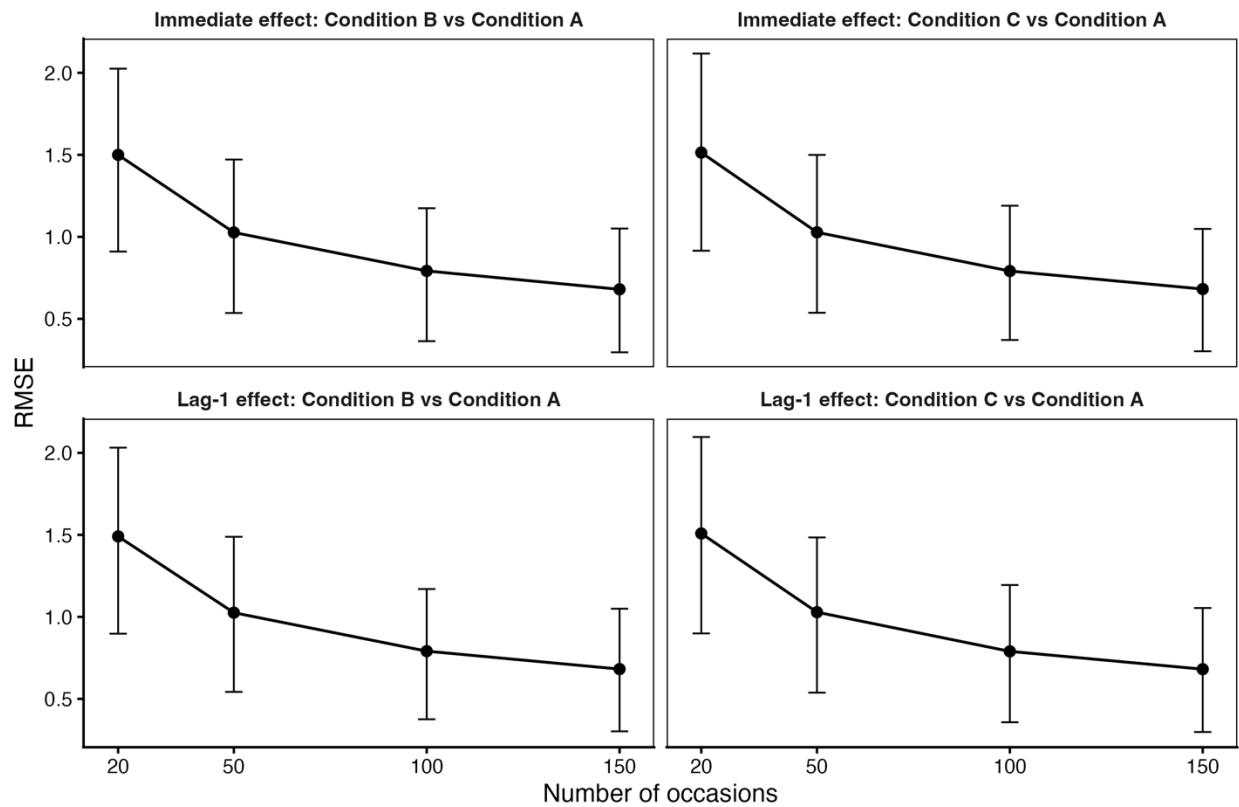
Relative Bias of the Three-Condition Regression-Based Estimator



RMSE showed the expected decrease as the number of occasions increased. Across the four causal contrasts, RMSE was largest in the shortest series and declined steadily as T increased. This pattern indicates that the estimator became more stable as more randomized occasions were available. RMSE was similar across the two active conditions and across immediate and lag-1 effects, suggesting that the proposed estimator behaved consistently across the vector-valued causal targets in the three-condition design. Figure 3 shows the RMSE across different number of occasions in the simulation.

Figure 3

RMSE of the Three-Condition Regression-Based Estimator

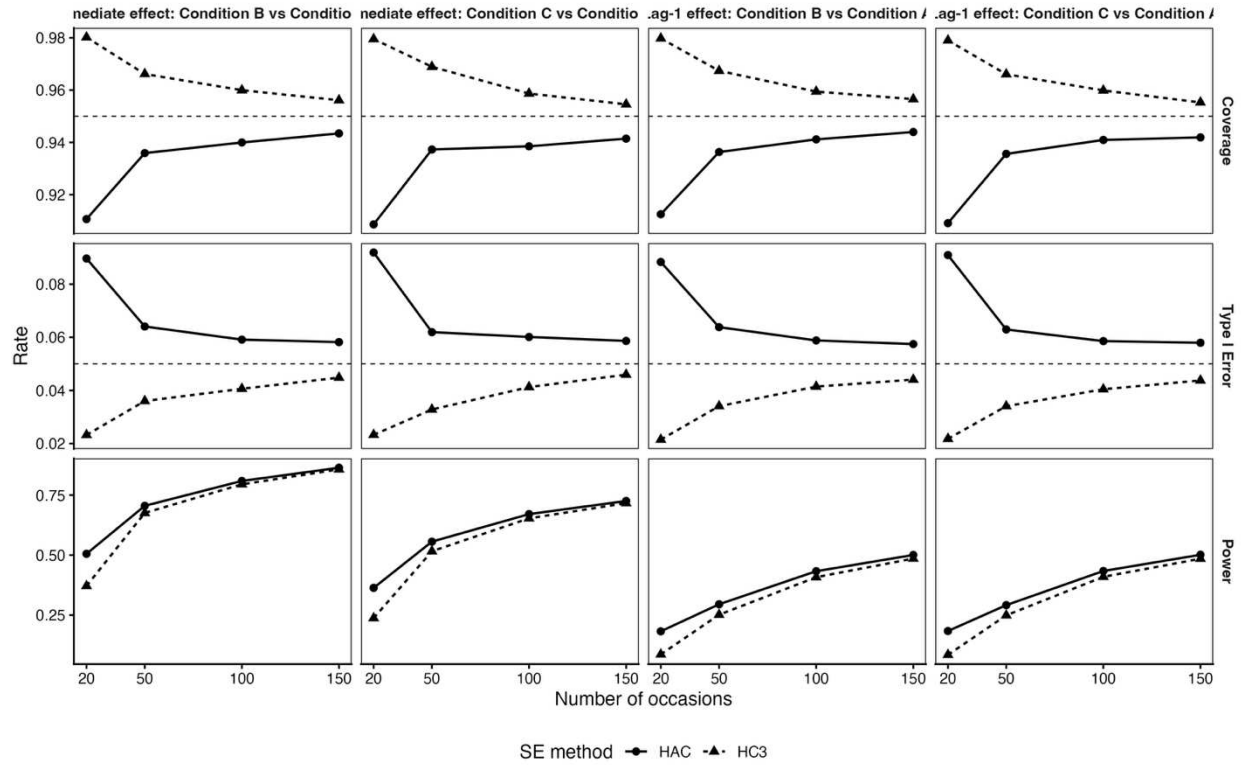


Inference performance depended on the standard error estimator, as shown in Figure 4.

HAC standard errors tended to produce slightly lower coverage and higher Type I error in shorter series, whereas HC3 standard errors were more conservative, especially when T was small. As the number of occasions increased, coverage and Type I error improved for both approaches. Power increased with series length and was higher for the larger immediate effects than for the smaller lag-1 effects.

Figure 4

Inference Performance of the Three-Condition Regression-Based Estimator



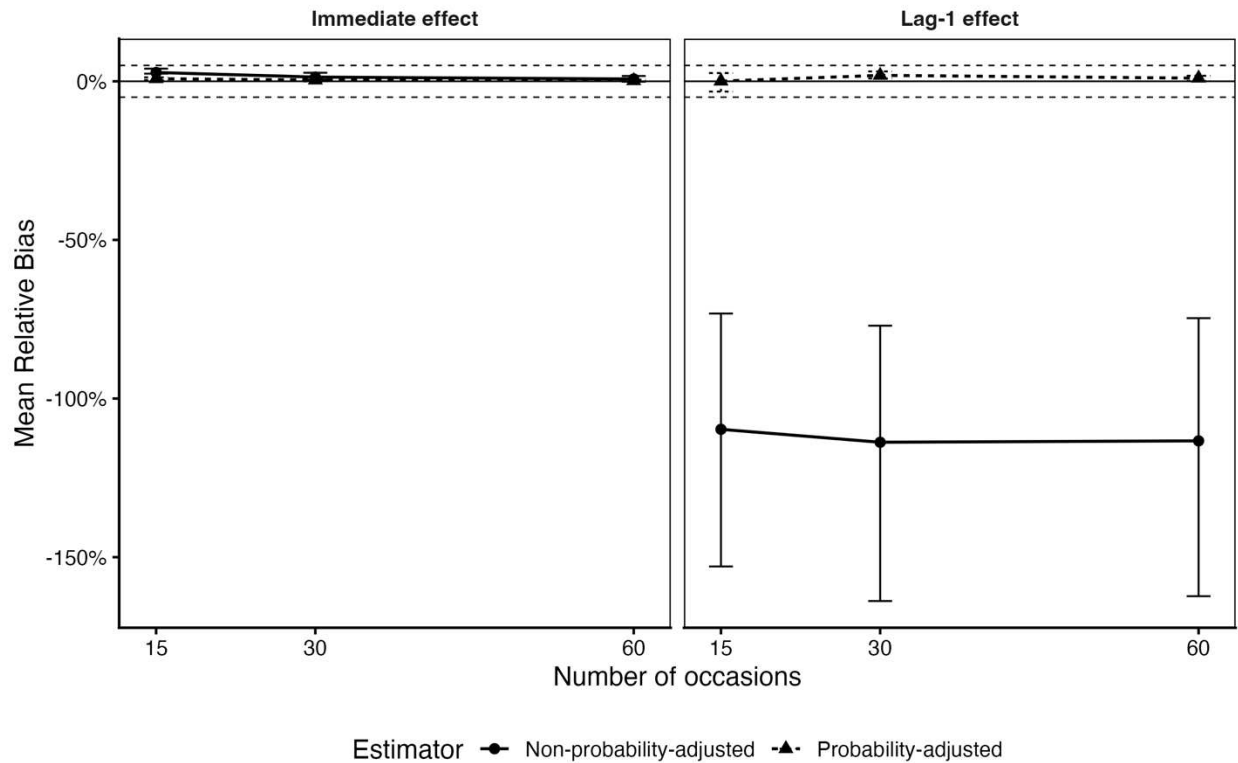
Simulation Study II

The relative bias results showed a clear difference between the non-probability-adjusted and probability-adjusted estimators. For the immediate effect, both estimators showed relative bias close to zero across series lengths. However, for the lag-1 effect, the non-probability-adjusted estimator showed substantial negative bias across all values of T . This bias did not disappear as the number of occasions increased, suggesting that the problem was not simply finite-sample variability but a systematic consequence of omitting the design-based probability adjustment. In contrast, the probability-adjusted estimator had relative bias close to zero for both the immediate and lag-1 effects. These results support the need to include p_t when the

randomization probability depends on prior treatment history. Figure 5 shows the mean relative bias for the immediate and lag-1 effects, comparing the non-probability-adjusted and probability-adjusted estimators.

Figure 5

Relative Bias of Restricted-Randomization Estimators

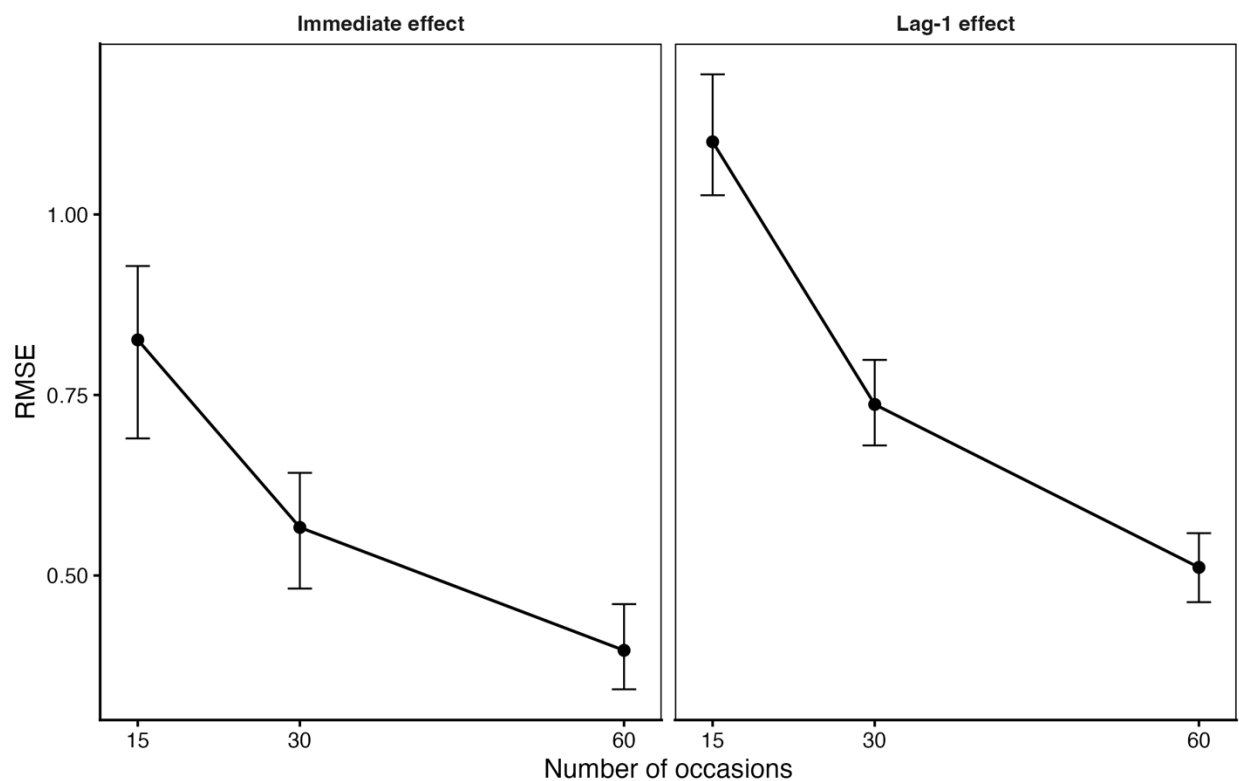


For the remaining results, we focused on the probability-adjusted estimator because the relative bias results showed that the non-probability-adjusted estimator was biased for the lag-1 effect. RMSE decreased as the number of occasions increased. For the immediate effect, RMSE was largest when $T = 15$, decreased at $T = 30$, and was smallest when $T = 60$. The same pattern was observed for the lag-1 effect, although RMSE was larger for the lag-1 effect than for the immediate effect across all series lengths. This result indicates that the probability-adjusted

estimator became more stable as more randomized occasions were available, but estimating lag-1 effects remained more difficult than estimating immediate effects in short restricted-randomization sequences. Figure 6 shows the RMSE of the probability-adjusted estimator across series lengths.

Figure 6

RMSE of the Probability-Adjusted Restricted-Randomization Estimator

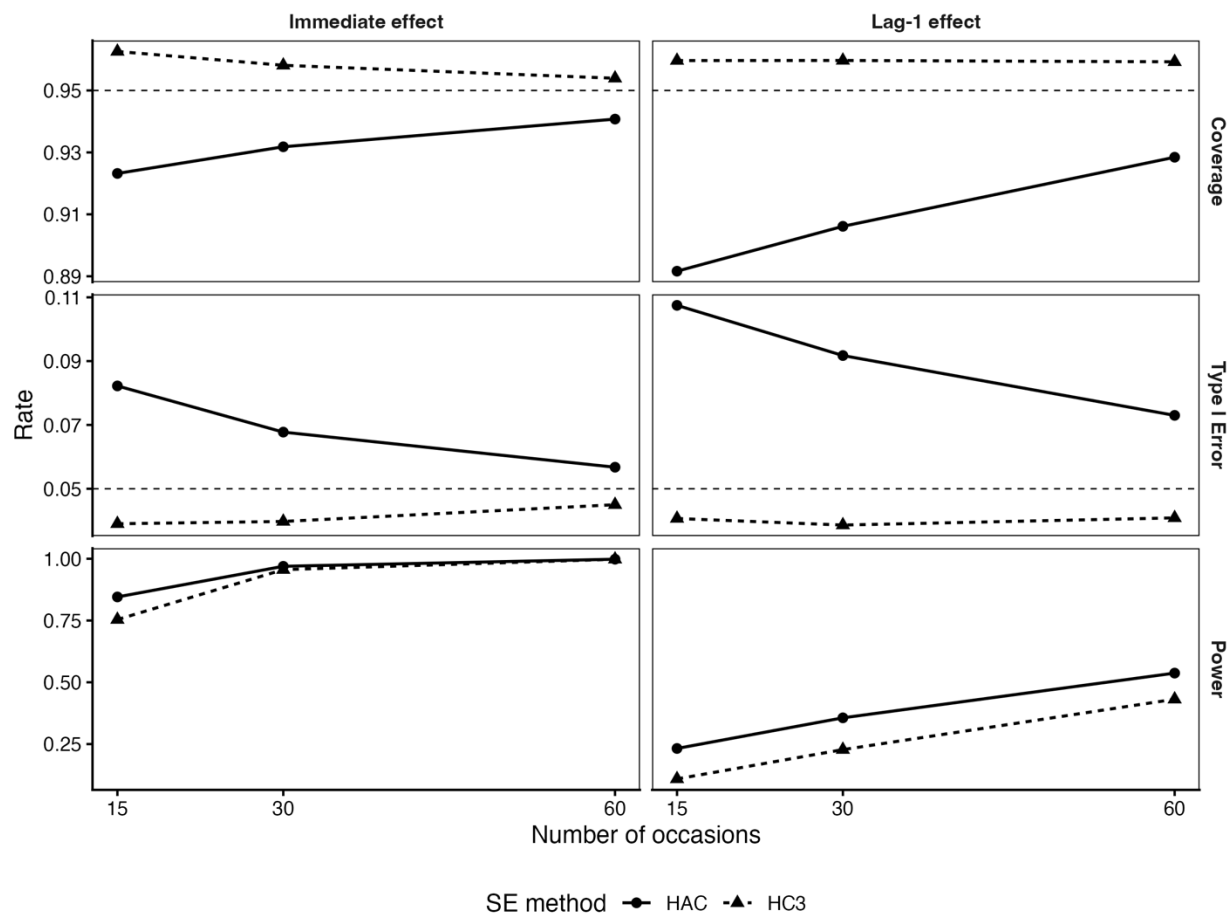


Inference performance for the probability-adjusted estimator depended on the standard error method, as shown in Figure 7. HAC standard errors produced lower coverage than HC3, especially for the lag-1 effect and when the series length was short. Coverage improved as T increased, but HAC coverage remained slightly below the nominal .95 level in several conditions. This pattern was also reflected in Type I error rates: HAC produced inflated Type I

error, particularly for the lag-1 effect in shorter series. HC3 standard errors were more conservative. They produced coverage closer to or slightly above the nominal level and lower Type I error rates across series lengths. However, this conservatism was associated with lower power, especially for the lag-1 effect. Power increased with T for both HAC and HC3, but power was consistently higher for the immediate effect than for the lag-1 effect. Overall, the probability-adjusted estimator showed good point-estimation performance, but inference remained challenging in short restricted-randomization series, especially for lag-1 effects.

Figure 7

Inference Performance of the Probability-Adjusted Restricted-Randomization Estimator



Empirical Demonstration

To illustrate how the proposed regression-based causal estimators can be implemented in applied psychological research, we applied the methods to the two empirical examples. These demonstrations show how the proposed estimators can be obtained from observed within-subject randomization design data and how the resulting coefficients can be interpreted.

Empirical Example I: Three-condition Intervention

The first empirical demonstration used data from the HeartSteps study (Klasnja et al., 2019), a micro-randomized trial designed to evaluate just-in-time activity suggestions. In the study, the same participant(s) were repeatedly randomized at different time points (known as decision points) when one of three conditions was assigned: no suggestion, a walking suggestion, or an anti-sedentary suggestion. The aim was to examine whether suggestions increased short-term physical activity. The outcome was the number of steps recorded in the 30 minutes after each decision point. The comparison condition was no suggestion, and the two treatment conditions were walking suggestions and anti-sedentary suggestions.

For the present demonstration, we selected one participant and included all eligible decision points at which the participant was randomized to one of the three conditions. We excluded decision points with missing outcome data, leaving 175 decision points in the final analysis. The three conditions were coded as follows: $W_t = 0$ for no suggestion, $W_t = 1$ for a walking suggestion, and $W_t = 2$ for an anti-sedentary suggestion. The known design probabilities were $p_{t0} = P(W_t = 0) = .40$, $p_{t1} = P(W_t = 1) = .30$, and $p_{t2} = P(W_t = 2) = .30$. Therefore, the treatment-indicator vector D_t included two indicators, one for a walking

suggestion and one for an anti-sedentary suggestion, with no suggestion serving as the comparison condition. With these probabilities, the assignment covariance matrix was

$$\Sigma = \begin{pmatrix} .21 & -.09 \\ -.09 & .21 \end{pmatrix}, \quad (35)$$

and its inverse was

$$\Sigma^{-1} = \begin{pmatrix} 5.83 & 2.50 \\ 2.50 & 5.83 \end{pmatrix}. \quad (36)$$

For each decision point, the current treatment-indicator vector D_t and its lag-1 version D_{t-1} were centered by their design expectations and weighted using Σ^{-1} . We then fit the working OLS regression with four transformed predictors: current walking suggestion, current anti-sedentary suggestion, lag-1 walking suggestion, and lag-1 anti-sedentary suggestion. The fitted coefficients were transformed back to the additive step-count scale using the same linear transformation described in Equation (22). Standard errors were computed using both HC3 and Newey–West HAC estimators.

The results are shown in Table 3. Using Newey-West HAC standard errors, the estimated immediate effect of a walking suggestion versus no suggestion was 129.04 additional steps in the 30 minutes after randomization, with a standard error of 77.45 and a 95% confidence interval from -23.84 to 281.92. The estimated immediate effect of an anti-sedentary suggestion versus no suggestion was 143.41 additional steps, with a standard error of 88.17 and a 95% confidence interval from -30.63 to 317.46. Although both immediate-effect estimates were positive, neither reached conventional statistical significance.

For the lag-1 effects, the estimated effect of a walking suggestion at the previous decision point was 54.42 steps, with a standard error of 87.70 and a 95% confidence interval from -118.69 to 227.53. The estimated lag-1 effect of an anti-sedentary suggestion was -95.60 steps, with a

standard error of 94.82 and a 95% confidence interval from -282.76 to 91.57. These estimates do not provide clear evidence that either type of suggestion had a residual effect on step counts at the next decision point.

Table 3

Regression-based causal estimates for one participant in the HeartSteps demonstration

Effect	SE type	Estimate	SE	95% CI
Immediate: walking vs no suggestion	HC3	129.04	83.73	[-36.24, 294.31]
Immediate: anti-sedentary vs no suggestion	HC3	143.41	97.37	[-48.79, 335.62]
Lag-1: walking vs no suggestion	HC3	54.42	88.21	[-119.69, 228.53]
Lag-1: anti-sedentary vs no suggestion	HC3	-95.60	91.33	[-275.87, 84.67]
Immediate: walking vs no suggestion	HAC	129.04	77.45	[-23.84, 281.92]
Immediate: anti-sedentary vs no suggestion	HAC	143.41	88.17	[-30.63, 317.46]
Lag-1: walking vs no suggestion	HAC	54.42	87.70	[-118.69, 227.53]
Lag-1: anti-sedentary vs no suggestion	HAC	-95.60	94.82	[-282.76, 91.57]

Empirical Example II: Restricted Randomization

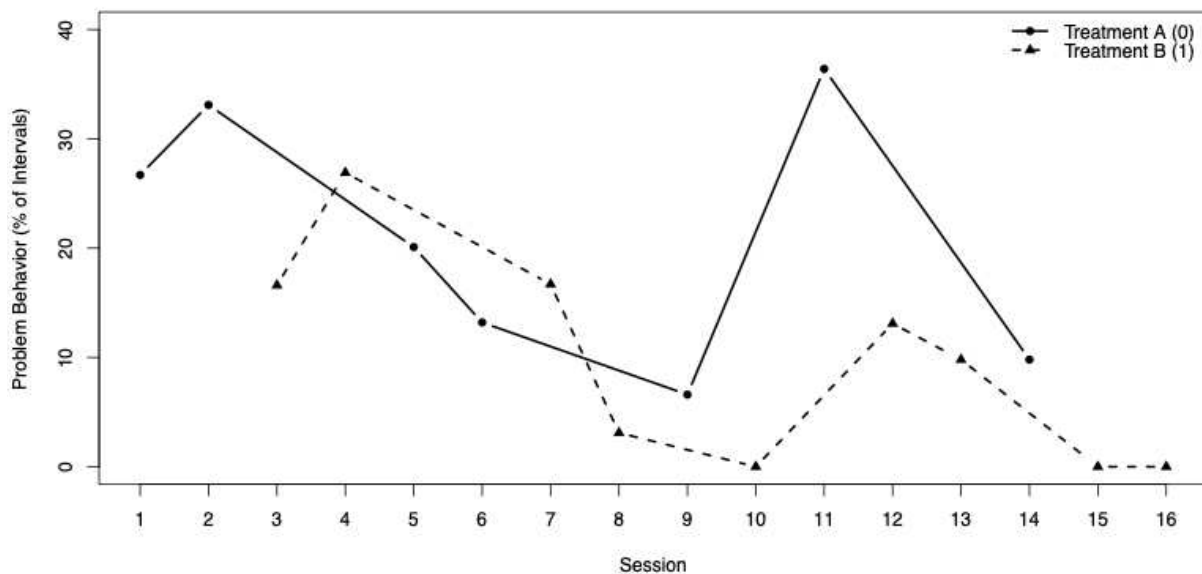
The second empirical demonstration used data from the randomized ATD design (Eilers & Hayes, 2015). The study compared a cognitive exercise with a control exercise for children with autism spectrum disorder. The goal was to examine whether the cognitive defusion

component could reduce problem behavior, particularly repetitive and restrictive behaviors. Two intervention conditions were alternated within each subject, and the treatment sequence was restricted so that no condition could be administered more than two times consecutively. The outcome was the percentage of intervals with problem behavior at each session.

For this demonstration, we focused on Jacob. The data were extracted from the published figure and consisted of 16 sessions shown in Figure 8. Treatment A and Treatment B were coded as $W_t = 0$ and $W_t = 1$, respectively. Under the no-more-than-two-consecutive-sessions restriction, 11 current-session assignments were randomized and 5 were deterministic.

Figure 8

Jacob's Problem Behavior Across Sessions by Treatment Condition in Eilers and Hayes (2015)



For each session, we computed the conditional probability $p_t = P(W_t = 1 \mid W_{1:t-1})$ implied by the restricted randomization rule and used it to construct the transformed current

treatment indicator $\tilde{W}_t$. On randomized sessions, $p_t = .50$ and $v_t=.25$, so the transformed indicator equals -2 for Treatment A and 2 for Treatment B, and it was set to zero when assignment was deterministic. Table 4 shows the values for Jacob's 16 sessions. The probability-adjusted working regression included $\tilde{W}_t$, $\tilde{W}_{t-1}$, and p_t as a design-based adjustment.

Table 4

Assignment Status and Transformed Treatment Indicator for Jacob

Session	Assignment status	Assigned condition	$\tilde{W}_t$
1	Randomized	Control (A)	-2
2	Randomized	Control (A)	-2
3	Deterministic	Treatment (B)	0
4	Randomized	Treatment (B)	2
5	Deterministic	Control (A)	0
6	Randomized	Control (A)	-2
7	Deterministic	Treatment (B)	0
8	Randomized	Treatment (B)	2
9	Deterministic	Control (A)	0
10	Randomized	Treatment (B)	2
11	Randomized	Control (A)	-2
12	Randomized	Treatment (B)	2
13	Randomized	Treatment (B)	2
14	Deterministic	Control (A)	0

Session	Assignment status	Assigned condition	$\tilde{W}_t$
15	Randomized	Treatment (B)	2
16	Randomized	Treatment (B)	2

Using the working OLS model in Equation (26), the estimated average immediate effect was -19.04 percentage points, with a standard error of 6.00 and a 95% confidence interval from -32.11 to -5.97. This suggests that Treatment B was associated with an immediate reduction in problem behavior relative to Treatment A for Jacob. The estimated average lag-1 effect was -2.81 percentage points, with a standard error of 5.62 and a 95% confidence interval from -15.06 to 9.44. Thus, there was limited evidence of a residual effect of the previous session's treatment on the current session's outcome after accounting for the current treatment and the restricted randomization mechanism.

Discussion

The present study introduced and advanced a regression-based causal inference framework for within-subject randomization designs in psychological research. The goal was to connect a familiar analytic tool (i.e., OLS regression) to design-based causal estimands defined under the potential-outcomes framework. The study focused on two treatment effects that are especially relevant for repeated within-subject experiments: the average immediate effect of the current treatment and the average lag-1 effect of the prior treatment. The study advanced existing methodology in two ways. First, it extended the regression-based causal estimator to multiple-arm intervention studies, where two or more active conditions are compared with a common reference condition. Second, it developed a probability-adjusted regression-based estimator for restricted randomization designs, where the treatment assignment probability depends on

treatment history with some sessions violating positivity assumption. The simulation studies provided evidence for the proposed approach. In the first simulation study, the proposed estimator recovered both immediate and lag-1 effects with small bias and robust inference with longer series length. In the restricted-randomization simulation, the probability-adjusted estimator also recovered both immediate and lag-1 effects with small bias. However, inference remained challenging in short restricted series: confidence intervals were undercovered and Type I error was inflated, especially for lag-1 effects. The empirical demonstrations illustrated how the proposed methods can be implemented in applied settings. In the HeartSteps study (Klasnja et al., 2019), the three-condition estimator produced interpretable estimates of additional steps following walking and anti-sedentary suggestions, as well as their lag-1 effects. In the restricted alternating-treatment example, the proposed probability-adjusted estimator separated the immediate effect from the lag-1 effect for Jacob. The immediate estimate suggested that Treatment B reduced problem behavior relative to Treatment A, whereas the lag-1 estimate did not provide clear evidence of residual treatment effects.

Relation to Existing Methods

The proposed framework differs from traditional analytic approaches used in within-subject designs. Traditional summaries, such as mean differences, nonoverlap indices, visual analysis, and standard regression summaries, often describe observed differences between conditions. These methods can be useful and accessible, but their causal interpretation may require additional assumptions about trends, comparability of sessions, and the absence of multiple treatment interference. In contrast, the proposed approach defines causal effects through potential outcomes and uses the known randomization mechanism to estimate those effects. The causal interpretation therefore comes from the design, not from assuming that the regression

model is the true data-generating process. The regression model is a working model for the design-based causal inference. Its purpose is not to correctly specify the outcome-generating process. Instead, it is a computational device for extracting randomized variation from the treatment sequence. A standard regression of the outcome on treatment indicators could be descriptive if it ignores the treatment assignment mechanism. The proposed regression centers and scales the treatment variables according to the known randomization design - this design-based transformation is what supports the causal interpretation.

The proposed framework also connects closely to recent design-based IPW and AIPW approaches for randomized alternating treatment designs. Luo and Li (2026) define average immediate and lag-1 causal effects and estimate them by weighting contrast-coded outcomes using the known assignment probabilities. The present study targets the same general problem: how to estimate causal effects in randomized within-subject designs when outcomes may depend on current and prior treatments. However, the implementation differs. Instead of asking applied researchers to construct weighted contrast outcomes directly, the present approach embeds the randomization information into transformed regression predictors. Its practical advantage is that it expresses this logic through regression, a tool already widely used by applied psychological researchers.

Limitations and Future Directions

First, the empirical HeartSteps demonstration used complete-timepoint analysis. In mobile health and intensive settings, missingness can arise from device non-wear, sensor failure, nonavailability, or participant behavior. If missingness is related to prior outcomes, context, or treatment assignment, complete-case analysis may change the target estimand or introduce bias. Future work should integrate and evaluate the proposed regression-based causal inference with

missing data strategies, such as multiple imputation or joint modeling of observation and outcome processes.

Second, the asymptotic theory for restricted within-subject randomization remains underdeveloped. Lin and Ding (2026)'s regression-based framework provides a foundation for time-series experiments, but the restricted-randomization designs common in psychological and behavioral applications introduce additional complications. Treatment probabilities may be history-dependent, deterministic occasions may violate positivity for some contrasts, and the lagged treatment predictor may be correlated with the current assignment mechanism. The simulation results suggest that the probability-adjusted estimator can reduce point-estimation bias under these conditions, but a full theoretical derivation is still needed to establish consistency, asymptotic normality, and valid variance estimation under realistic restricted designs.

Third, inference remained difficult in short restricted series. In Simulation Study II, the proposed estimator had small bias, but HAC confidence intervals undercovered and Type I error rates were inflated, particularly for lag-1 effects. On the other hand, HC3 controlled Type I error more closely, although its confidence intervals were conservative and its power was lower. Because many alternating treatment designs have short series, standard Wald-type inference should be interpreted cautiously. In particular, a nonsignificant lag-1 estimate should not be taken as evidence of no carryover effect, especially when few occasions provide randomized lag-1 contrasts.

Fourth, the present study focused on a single experimental unit at a time. It did not assess whether the average immediate and lag-1 effects vary across units in studies that include multiple participants, classrooms, dyads, or settings. Future research could extend the framework to

treatment-effect heterogeneity across experimental units. Moment-based methods could be developed to estimate variation in the average immediate and lag-1 effects across units while accounting for estimation error within each unit.

Reference

- Bailey, D. H., Jung, A. J., Beltz, A. M., Eronen, M. I., Gische, C., Hamaker, E. L., Kording, K. P., Lebel, C., Lindquist, M. A., Moeller, J., Razi, A., Rohrer, J. M., Zhang, B., & Murayama, K. (2024). Causal inference on human behaviour. *Nature Human Behaviour*, 8(8), 1448–1459. <https://doi.org/10.1038/s41562-024-01939-z>
- Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, 12(2), 199–210. <https://doi.org/10.1901/jaba.1979.12-199>
- Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, 4(1), 19–21. <https://doi.org/10.1111/j.1745-6924.2009.01088.x>
- Bojinov, I., & Shephard, N. (2019). Time series experiments and causal estimands: Exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528), 1665–1682. <https://doi.org/10.1080/01621459.2018.1527225>
- Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, 5(8), 980–989. <https://doi.org/10.1038/s41562-021-01143-3>

- Davidson, K. W., Peacock, J., Kronish, I. M., & Edmondson, D. (2014). Personalizing behavioral interventions through single-patient (N-of-1) trials. *Social and Personality Psychology Compass*, 8(8), 408–421. <https://doi.org/10.1111/spc3.12121>
- Daza, E. J. (2018). Causal analysis of self-tracked time series data using a counterfactual framework for N-of-1 trials. *Methods of Information in Medicine*, 57(S 01), e10–e21. <https://doi.org/10.3414/ME16-02-0044>
- De Carvalho, M. C. L., de Matos Dourado Simões, M., Adler, U. C., Junior, A. B. V., de Sousa, C. N. S., & Sanders, L. L. O. (2026). N-of-1 trials in clinical research: Methodological foundations, statistical approaches and implementation challenges. *British Journal of Clinical Pharmacology*, 92(3), 809–821. <https://doi.org/10.1002/bcp.70382>
- Edgington, E. S. (1980). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment*, 2, 19–28.
- Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy*, 34(7), 567–574. [https://doi.org/10.1016/0005-7967\(96\)00012-5](https://doi.org/10.1016/0005-7967(96)00012-5)
- Eilers, H. J., & Hayes, S. C. (2015). Exposure and response prevention therapy with cognitive defusion exercises to reduce repetitive and restrictive behaviors displayed by children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 19, 18–31. <https://doi.org/10.1016/j.rasd.2014.12.014>
- Epstein, L. H., & Dallery, J. (2022). The family of single-case experimental designs. *Harvard Data Science Review*, 4(SI3). <https://doi.org/10.1162/99608f92.ff9300a8>

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Prentice Hall.

Hekler, E. B., Klasnja, P., Chevance, G., Golaszewski, N. M., Lewis, D., & Sim, I. (2019). Why we need a small data paradigm. *BMC Medicine*, 17, Article 133.
<https://doi.org/10.1186/s12916-019-1366-x>

Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, 3(1), 51–64. <https://doi.org/10.1016/j.jcbs.2013.10.002>

Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single-subject research: Description of designs and discussion of problems. *Topics in Early Childhood Special Education*, 14(1), 119–145. <https://doi.org/10.1177/027112149401400111>

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960. <https://doi.org/10.1080/01621459.1986.10478354>

Kazdin, A. E. (2021). *Single-case research designs: Methods for clinical and applied settings* (3rd ed.). Oxford University Press.

Kim, J.-S. (2010). Within-subjects design. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1639–1644). SAGE Publications.
<https://doi.org/10.4135/9781412961288.n503>

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time

adaptive interventions. *Health Psychology*, 34S, 1220–1228.

<https://doi.org/10.1037/hea0000305>

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., & Murphy, S. A. (2019). Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6), 573–582. <https://doi.org/10.1093/abm/kay067>

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, 15(2), 124–144. <https://doi.org/10.1037/a0017736>

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine*, 8(2), 161–173. <https://doi.org/10.2217/pme.11.7>

Lin, Z., & Ding, P. (2025). *Unifying regression-based and design-based causal inference in time-series experiments*. arXiv. <https://arxiv.org/abs/2510.22864>

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3), 217–224. <https://doi.org/10.1080/00031305.2000.10474549>

Luo, W., & Li, C. (2026). *Defining and estimating causal effects in randomized alternating treatment design for single-case experiments: A counterfactual approach*. Research Square. <https://doi.org/10.21203/rs.3.rs-9237753/v1>

- Manolov, R., & Onghena, P. (2018). Analyzing data from single-case alternating treatments designs. *Psychological Methods*, 23(3), 480–504. <https://doi.org/10.1037/met0000133>
- McDonald, A. M., Knight, R. C., Campbell, M. K., Entwistle, V. A., Grant, A. M., Cook, J. A., Elbourne, D. R., Francis, D., Garcia, J., Roberts, I., & Snowdon, C. (2006). What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials*, 7, Article 9. <https://doi.org/10.1186/1745-6215-7-9>
- Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, 2(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708. <https://doi.org/10.2307/1913610>
- Qian, T., Walton, A. E., Collins, L. M., Klasnja, P., Lanza, S. T., Nahum-Shani, I., Rabbi, M., Russell, M. A., Walton, M. A., Yoo, H., & Murphy, S. A. (2022). The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychological Methods*, 27(5), 874–894. <https://doi.org/10.1037/met0000283>
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

<https://doi.org/10.1037/h0037350>

Shaffer, J. A., Kronish, I. M., Falzon, L., Cheung, Y. K., & Davidson, K. W. (2018). N-of-1 randomized intervention trials in health psychology: A systematic review and methodology critique. *Annals of Behavioral Medicine*, 52(9), 731–742.

<https://doi.org/10.1093/abm/kax026>

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Nikles, J., Tate, R., Johnston, B. C., Zucker, D., Shadish, W. R., Kravitz, R., Guyatt, G., Altman, D. G., Moher, D., & Vohra, S. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration. *BMJ*, 350, Article h1793. <https://doi.org/10.1136/bmj.h1793>

Sully, B. G. O., Julious, S. A., & Nicholl, J. (2013). A reinvestigation of recruitment to randomised, controlled, multicenter trials: A review of trials funded by two UK funding agencies. *Trials*, 14, Article 166. <https://doi.org/10.1186/1745-6215-14-166>

Walters, S. J., Bonacho dos Anjos Henriques-Cadby, I., Bortolami, O., Flight, L., Hind, D., Jacques, R. M., Knox, C., Nadin, B., Rothwell, J., Surtees, M., & Julious, S. A. (2017). Recruitment and retention of participants in randomised controlled trials: A review of trials funded and published by the United Kingdom Health Technology Assessment

Programme. *BMJ Open*, 7(3), Article e015276. <https://doi.org/10.1136/bmjopen-2016-015276>

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. Institute of Education Sciences, U.S. Department of Education. <https://ies.ed.gov/ncee/wwc/Handbooks>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementalmaterialssubmission.docx](#)