APPENDIX
Multi-cell experiments for marginal treatment effect estimation of digital ads*

Published in *Management Science*

Caio Waisman Brett R. Gordon
Kellogg School of Management
Northwestern University

July 19, 2024

* E-mail addresses for correspondence: caio.waisman@kellogg.northwestern.edu, b-gordon@kellogg.northwestern.edu.

Appendix

Table of Contents

A Rewriting the firm's decision problem

We demonstrate how the firm's decision problem given in equation (2) can be rewritten as in equation (7). We begin by deriving the terms in equation (6). First, we have that:

$$\begin{aligned}
\mathbb{E}\left[Y_1|U \leq \phi\right] &= \int_0^{\phi} \int_{y_1 \in \mathcal{Y}_1} y_1 \frac{f(y_1, u)}{\Pr(U \leq \phi)} du dy_1 \\
&= \int_0^{\phi} \left(\int_{y_1 \in \mathcal{Y}_1} y_1 f(y_1|u) dy_1 \right) \frac{f(u)}{\phi} du \\
&= \int_0^{\phi} \mathbb{E}\left[Y_1|U = u\right] \times \frac{1}{\phi} du \\
&\equiv \int_0^{\phi} m_1(u) \frac{1}{\phi} du, \qquad \text{(A.1)}
\end{aligned}$$

where in the third equality we used that $U \sim U(0,1)$ and defined $m_1(u) \equiv \mathbb{E}\left[Y_1|U = u\right]$. Second,

$$\begin{aligned}
\mathbb{E}\left[Y_0|U > \phi\right] &= \int_{\phi}^{1} \int_{y_0 \in \mathcal{Y}_0} y_0 \frac{f(y_0, u)}{\Pr(U > \phi)} du dy_0 \\
&= \int_{\phi}^{1} \left(\int_{y_0 \in \mathcal{Y}_0} y_0 f(y_0|u) dy_0 \right) \frac{f(u)}{1-\phi} du \\
&= \int_{\phi}^{1} \mathbb{E}\left[Y_0|U = u\right] \times \frac{1}{1-\phi} du \\
&\equiv \int_{\phi}^{1} m_0(u) \frac{1}{1-\phi} du, \qquad \text{(A.2)}
\end{aligned}$$

where in the third equality we also used that $U \sim U(0,1)$ and defined $m_0(u) \equiv \mathbb{E}\left[Y_0|U = u\right]$.

Plugging equations (A.1) and (A.2) back into equation (2) yields:

$$\begin{aligned}
&\max_{\phi \in [0,1]} \left(\delta \times \left\{ \int_0^{\phi} m_1(u) du + \int_{\phi}^{1} m_0(u) du \right\} - \kappa(\phi) \right) \\
&\max_{\phi \in [0,1]} \left(\delta \times \int_0^{1} m_0(u) du + \delta \times \left\{ \int_0^{\phi} \left[m_1(u) - m_0(u) \right] du \right\} - \kappa(\phi) \right) \\
&\max_{\phi \in [0,1]} \left(\delta \times \mathbb{E}\left[Y_0\right] + \delta \times \int_0^{\phi} \text{MTE}(u) du - \kappa(\phi) \right) \\
&\max_{\phi \in [0,1]} \left(\delta \times \int_0^{\phi} \text{MTE}(u) du - \kappa(\phi) \right),
\end{aligned}$$

which establishes equation (7).

B Details of calibrated simulations

Below we provide more details about how we simulate data based on Study 4 from Gordon et al. (2019) for our analysis in Section 4.

1. **MTR functions specification.** We specify true MTR functions of the form:[1]

$$m_1(u) = m_{10} + m_{11}u + m_{12}u^2$$
$$m_0(u) = m_{00} + m_{01}u + m_{02}u^2 + m_{03}u^3$$

   We chose these functional forms because they are the polynomials of lowest degree that the simplest version of our design—with only two cells—cannot recover. With three or more cells, our approach can perfectly recover the true MTR functions. These forms imply that the MTE is a cubic polynomial.

2. **Implied moments.** We need to choose parameter values such that the implied moments match the observed ψ_{11}, ψ_{01} and ψ_{00}. Under these functional forms, from equations (10), (11), and (12) it follows that:

$$\psi_{11} = m_{10} + \frac{m_{11}}{2}p(1) + \frac{m_{12}}{3}p(1)^2$$
$$\psi_{01} = m_{00} + \frac{m_{01}}{2}\left[1 + p(1)\right] + \frac{m_{02}}{3}\left[1 + p(1) + p(1)^2\right] + \frac{m_{03}}{4}\left[1 + p(1) + p(1)^2 + p(1)^3\right]$$
$$\psi_{00} = m_{00} + \frac{m_{01}}{2} + \frac{m_{02}}{3} + \frac{m_{03}}{4}$$

3. **Choosing parameter values.** Based on the observed moments, we need to choose seven parameters, $\left\{\{m_{1c}\}_{c=0}^{2}, \{m_{0c}\}_{c=0}^{3}\right\}$, to satisfy the three constraints above, plus the additional constraint that the MTR functions must be between 0 and 1, because the outcome of Study 4 is binary (purchase). We consider two sets of parameters to illustrate our proposed approach, labeled DGP1 and DGP2. For each DGP, we choose four parameters and then solve for the remaining three by plugging the observed ψ_{11}, ψ_{01}, ψ_{00}, and $p(1)$ into the system of equations above. We choose different set of values for the DGPs to generate different patterns of the MTE function. Since both DGPs correspond to functions that are very close to quadratic, we expect that a two-cell design will suffice to obtain good approximations.

[1] Here we follow the approach from Mogstad et al. (2018), who directly specified the MTR functions in their simulations. An alternative and commonly used approach, employed, for instance, in Heckman and Vytlacil (2007), is to specify the joint distribution of Y_1, Y_0 and U, which is often normal, and work with the implied MTR functions. We follow this alternative approach in Appendix C, where we consider the role of multiple ad exposures.

4. **Simulating the multi-cell experiment.** We use these MTE functions to simulate data from our experimental design with two cells. This design limits us to a linear approximation to $m_1(\cdot)$ and a quadratic approximation to $m_0(\cdot)$. First, we set the propensity scores in our simulated multi-cell experiment. Because Study 4 assigned users to be eligible to receive treatment with probability 0.7, we consider this to be one cell and add a second cell in which this probability equals 0.3. For the first cell we keep the propensity score at the original value, $p(Z_1 = 1) = 0.37$, and set the propensity score for the second cell so that $\Pr(Z_1 = 1|C = 1) \times p(Z_1 = 1) = \Pr(Z_2 = 1|C = 2) \times p(Z_2 = 1)$, implying that $p(Z_2 = 1) \approx 0.863$. We use these propensity scores and the underlying MTE functions to compute the ψs from equations (10) and (11) that would have been observed had our design been implemented.

C Dealing with multiple ad exposures

C.1 General incorporation of observable characteristics

We now demonstrate how to incorporate observable characteristics, captured in a vector X, into the model above. Three changes have to be made. First, equation (3) is replaced with:

$$D = \mathbb{1}\{p(Z, X) \geq U\}. \tag{C.1}$$

Importantly, notice that the error term, U, remains additively separable.

Second, Assumption 1 is replaced with:

Assumption C.1.
(i) $U \perp\!\!\!\perp Z|X$, *where* $\perp\!\!\!\perp$ *denotes statistical independence.*
(ii) $\mathbb{E}[Y_d|Z, X, U] = \mathbb{E}[Y_d|X, U]$ *and* $\mathbb{E}\left[Y_d^2\right] < \infty$ *for* $d \in \{0, 1\}$.
(iii) U *is continuously distributed conditional on* X.

Finally, we replace Assumption 2 with:

Assumption C.2.
(i) $\Pr(Z_c = z|X, \mathcal{C} = c) \in (0, 1)$ *for all* X, c *and all* z.
(ii) $p(Z_c = 1|X) \equiv \Pr(D = 1|X, Z_c = 1) \in (0, 1)$ *for all* c *and* X.
(iii) $p(Z_c = 1|X) \neq p(Z_{c'} = 1|X)$ *for all* $c \neq c'$ *and* X.

In summary, Assumptions C.1 and C.2 simply add conditioning on X to Assumptions 1

and 2. The MTR and MTE functions become $m_d(u,x) \equiv \mathbb{E}[Y_d|U=u, X=x]$, where $d \in \{0,1\}$, and $\text{MTE}(u,x) \equiv \mathbb{E}[Y_1 - Y_0|X=x, U=u] = m_1(x,u) - m_0(x,u)$, respectively.

C.2 Number of previous ad exposures as a covariate

In certain cases, one can expect that multiple exposures to an ad can have a significant impact on outcomes. We now show how multiple ad impression opportunities can be incorporated as a covariate into our setup.

To this end, assume that each user can be exposed to the ad S times. We index all variables by $s = 1, \ldots, S$, to indicate that they are associated with impression opportunity s. We define X_s as the number of times the user was exposed to the ad *prior* to impression opportunity s.

In this specific case in which previous exposures is a covariate, we use the budget designated to each user as the instrument that generates variation in the probability of exposures and in the number of previous ad exposures. We denote the budget originally designated to a user by B. As the advertiser shows their ad over impression opportunities, this budget diminishes due to the auction payments. We denote the remaining available budget at impression opportunity s by B_s.

We follow the model with essential heterogeneity from Heckman et al. (2006):[2]

$$\begin{aligned} Y_{1s} &= \mu_1(X_s) + \epsilon_{1s} \\ Y_{0s} &= \mu_0(X_s) + \epsilon_{0s} \\ D_s &= \mathbb{1}\{p(B_s, X_s) \geq U_s\}. \end{aligned} \tag{C.2}$$

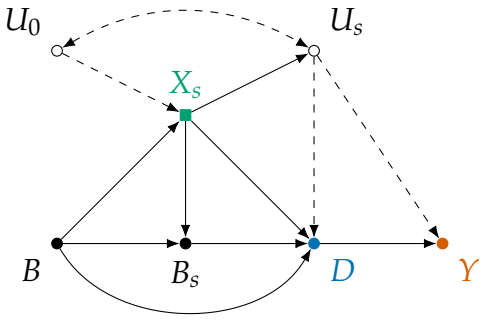
We complete this model with the following assumption. As we discuss below, this assumption is key for the analysis we describe next.

Assumption C.3.
Conditional on X_s, ϵ_{1s}, ϵ_{0s}, and U_s are i.i.d. across s.

The i.i.d. condition across s deserves attention, especially regarding U_s. We interpret the variable U as the ease with which a user can be exposed to the ad. Thus, it would be reasonable to expect that this term would be correlated across s. In other words, users may have a persistent unobservable that determines how easy it is to expose them to an

[2]Conceptually, this model could be further generalized by letting $Y_d = \mu_d(X, \epsilon_d)$, where $d \in \{0,1\}$. This may be desirable to explicitly model, for example, binary potential outcomes. However, we maintain the additively separable structure for simplicity.

Figure C.1: DAG showing violation of Assumption C.3



ad. Consequently, Assumption C.3 maintains that conditioning on the X_s eliminates this persistent term. The DAG presented in Figure C.1 illustrates a violation of Assumption C.3.[3]

In Figure C.1, Assumption C.3 is violated: even conditional on X_s, the unobservables U_s remain correlated across s. As a result, the available budget, B_s, becomes an invalid instrument for D_s. This could happen if among people exposed to many ads (high X_s), users who are hard to reach (high U_s) had a high initial budget (large B) relative to users who are easier to reach (low U_s). This would invalidate the available budget, B_s, as an instrument for D_s.

Assumption C.3 restores this validity by effectively assuming away the dependence of X_s on U_s. This arguably is a strong assumption, which could become more palatable if additional variables, such as user and impression opportunity characteristics, were incorporated into X. However, this is a caveat that should be taken into consideration.

C.3 Budget optimization problem with multiple exposures

We now present the advertiser's optimization problem in the presence of multiple ad exposures, which can be seen as a generalized version of the problem presented in Section 2.3. Unlike the original problem, it will be more convenient to cast it as a function of B. Hence, the problem is given by:

$$\max_{B \in \mathbb{R}_+} \delta \times \mathbf{E}\left[Y|B\right] - B \tag{C.3}$$

[3]We thank an anonymous reviewer for providing us with this figure.

where all variables are as previously defined. Notice that:

$$\begin{aligned}
\mathbf{E}\left[Y|B\right] &= \mathbf{E}\left\{\mathbf{E}\left[Y|X,B\right]|B\right\} \\
&= \sum_{x=0}^{S-1} \Pr\left(X=x|B\right) \times \mathbf{E}\left[Y|X=x,B\right] \\
&= \sum_{x=0}^{S-1} \Pr\left(X=x|B\right) \times \left\{\mathbf{E}\left[DY_1|X=x,B\right] + \mathbf{E}\left[(1-D)Y_0|X=x,B\right]\right\} \\
&= \sum_{x=0}^{S-1} \Pr\left(X=x|B\right) \times \left\{\int_0^{p(B,x)} m_1(x,u)du + \int_{p(B,x)}^1 m_0(x,u)du\right\}. \qquad \text{(C.4)}
\end{aligned}$$

Note that we are slightly abusing notation by denoting the propensity score as a function of the original budget, B, instead of the available budget after x impressions. Conditional on $x > 0$, we expect that the available budget to be lower than B but this difference depends on the particular ad payment structure. We will illustrate this in our simulations below.

Plugging (C.4) into (C.3) yields:

$$\max_{B\in\mathbb{R}_+} \left\{\delta \times \sum_{x=0}^{S-1} \Pr\left(X=x|B\right) \times \left[\int_0^{p(B,x)} m_1(x,u)du + \int_{p(B,x)}^1 m_0(x,u)du\right] - B\right\}. \qquad \text{(C.5)}$$

Whether there is a unique solution to this problem depends on the functions $\Pr\left(X=x|B\right)$ and $p(B,x)$, which are either known (i.e., the platform reports them to the advertiser) or estimable since X, B, and D are observable. Hence, the remaining pieces that have to be estimated in order for the advertiser to solve this budget optimization problem are the MTR functions. We now discuss how to accomplish this estimation task.

C.4 Sequential approximations to MTE function

We propose a sequential approach that approximates the MTE function at each value of X_s separately. This is motivated by the fact that: (i) online advertising experiments involve a large number of users; (ii) the data are often stored sequentially as they arrive for the duration of the experiment; and (iii) the only variable we are including in X_s is the previous number of impressions. Should other variables be included in the model, alternative approaches can also be considered.

Under Assumption C.3, the MTR functions take the form:

$$m_d(x,u) = \mu_d(x) + \mathbb{E}\left[\epsilon_{ds} | X_s = x, U_s = u\right], \tag{C.6}$$

where $d \in \{0,1\}$. Hence, using a multi-cell experimental design with sequential storage of data allows us to recover the following moments:

$$\begin{aligned}
\psi_{1bx} &= \mathbb{E}\left[Y | D_s = 1, X_s = x, B_s = b\right] \\
&= \mathbb{E}\left[Y_1 | U_s \leq p(b,x), X_s = x, B_s = b\right] \\
&= \mu_1(x) + \frac{1}{p(b,x)} \int_0^{p(b,x)} \mathbb{E}\left[\epsilon_{1s} | U_s = u\right] du
\end{aligned} \tag{C.7}$$

and

$$\begin{aligned}
\psi_{0bx} &= \mathbb{E}\left[Y | D_s = 0, X_s = x, B_s = b\right] \\
&= \mathbb{E}\left[Y_1 | U_s > p(b,x), X_s = x, B_s = b\right] \\
&= \mu_0(x) + \frac{1}{1 - p(b,x)} \int_{p(b,x)}^1 \mathbb{E}\left[\epsilon_{0s} | U_s = u\right] du.
\end{aligned} \tag{C.8}$$

We can approximate the MTE function at each value of x separately using ψ_{1bx}, ψ_{0bx}, and $p(b,x)$ following the procedure described in Section 3.3.

This procedure has an important feature. As impressions are obtained, the available budget decreases; that is, the number of different values of B_s decreases as X_s increases. Consequently, the degree of the approximating polynomial is lower for higher values of previous exposures to the ad, rendering this approximation, at least theoretically, poorer.

C.5 Calibrated simulations with multiple exposures

We now describe the specification we chose for our calibrated simulations with multiple ad impression opportunities. This specification is motivated solely for simplicity, but the parameter values are chosen based on the data from the Facebook experiment we consider.

C.5.1 Distributional and parametric assumptions

To implement our calibrated simulations, we make the following normality assumption:

$$\begin{bmatrix} \epsilon_{1s} \\ \epsilon_{0s} \\ V_s \end{bmatrix} \Big| X_s \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 & \rho_{1V}\sigma_1 \\ \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 & \rho_{0V}\sigma_0 \\ \rho_{1V}\sigma_1 & \rho_{0V}\sigma_0 & 1 \end{bmatrix} \right) \tag{C.9}$$

where $V_s = \Phi^{-1}(U_s)$ and where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. We further assume that:

$$\mu_d(x) = \mu_{d0} + \mu_{d1}x + \mu_{d2}x^2 \tag{C.10}$$

where $d \in \{0,1\}$. Consequently, we obtain:

$$m_d(x,u) = \mu_{d0} + \mu_{d1}x + \mu_{d2}x^2 + \rho_{dV}\sigma_d\Phi^{-1}(u). \tag{C.11}$$

The normality assumption in (C.9) is oftentimes maintained and is of great aid in identifying and estimating the MTE function. As discussed, for example, in Mogstad and Torgovitsky (2018), all that is necessary to recover $\rho_{dV}\sigma_d$, where $d \in \{0,1\}$, is for the propensity score to have two distinct values between 0 and 1 at just one value of x. Consequently, with enough variation in x to estimate $\mu_d(\cdot)$ it is possible to recover the entire MTR functions under this weaker condition.

However, this convenience also imposes non-trivial restrictions on the model. For instance, it imposes that the MTR functions are monotonic in u, ruling out the DGPs we considered in Section 4, and that they tend to plus and minus infinity at the extremes of the unit interval, as depicted in Figure C.2. On the other hand, the approach of approximating the MTR functions with polynomials from BMW is agnostic and does not impose such restrictions on the DGP.

C.5.2 Features of simulated experiment

First, we set $S = 3$. In practice, the number of potential ad impressions can be much larger and is a function of the duration of the experiment. We choose a low value solely for simplicity. In addition, such reduction is sometimes implemented in practice via bracketing. For instance, $s = 1$ can correspond to the first five impressions, $s = 2$ to the following five, and so on.

Second, we assume that:

$$p(B_s, X_s) = \frac{B_s}{B_s + t}, \tag{C.12}$$

where $t > 0$. This is a simple way of ensuring that the probability of exposing a user to the ad is zero when there is no budget left, that is, when $B_s = 0$. We choose the parameter t to generate variation in the propensity score across the different values of B_s. Also, notice that this probability does not depend on X_s, a choice we made solely for simplicity. X_s could be easily incorporated into this propensity function, for instance, via frequency caps.

Third, we assume that the experiment randomizes users into five groups determined by the initial budget allocated to them. In particular, we set $B \in \{0, 1, 2, 3, 4\}$.

Finally, when the advertiser wins an impression opportunity, they make a payment and the available budget decreases. We simply assume that:

$$B_{s+1} = \begin{cases} B_s - 1 & \text{if } D_s = 1 \\ B_s & \text{if } D_s = 0 \end{cases} \tag{C.13}$$

We choose the budget to decrease by one to facilitate the simulation by aligning the evolution of budget with the initial budget allocation. In practice, decreases in the available budget are determined by auction payments, which almost always vary across impression opportunities.

C.5.3 Choice of parameter values

Under this setup, we have to choose values for ten parameters: μ_{10}, μ_{11}, μ_{12}, μ_{00}, μ_{01}, μ_{02}, σ_1, σ_0, ρ_{1V}, and ρ_{0V}.[4] To do so, we use data from the same Study 4 we considered in Section 4. It is important to recall that Study 4 only randomized eligibility to ad exposure, denoted by the dummy variable Z, and not budget levels.

We set $\sigma_d^2 = \text{Var}(Y|Z = d)$, where $d \in \{0, 1\}$, and obtain $\sigma_1 = 0.0185584$ and $\sigma_0 = 0.0158245$. Nevertheless, it is important to note that, in practice, these equalities are unlikely to hold.

The data to which we have access to calibrate $\mu_d \equiv [\mu_{d0}, \mu_{d1}, \mu_{d2}]'$, where $d \in \{0, 1\}$, consist of the following moments: $\mathbb{E}[Y|D = 0, Z = 0]$, $\mathbb{E}[Y|D = 0, Z = 1]$, $\mathbb{E}[Y|X = 1]$,

[4] Notice that the parameter ρ_{10} is irrelevant for the purposes of computing the MTE function.

$\mathbb{E}[Y|X=1]$, $\mathbb{E}[Y|X=2]$, and $\mathbb{E}[Y|X=3]$.[5] These expressions and their interpretation warrant a few comments.

We only have access to aggregate results at the end of the experiment, but not to the sequential evolution of results. Thus, Y and X are the observed outcomes and total number of ad exposures, respectively, at the end of the experiment. As a consequence, we omit the conditioning on X when $D = 0$ because, by construction, it is equal to zero. Furthermore, also by construction, $X > 0$ if and only if $D = Z = 1$, so we omit the conditioning on $D = 1$ and $Z = 1$ when $X > 0$. Given this structure, we only have two moments informative of $m_0(\cdot)$, $\mathbb{E}[Y|D=0, Z=0]$ and $\mathbb{E}[Y|D=0, Z=1]$, and three moments informative of $m_1(\cdot)$, $\mathbb{E}[Y|X=1]$, $\mathbb{E}[Y|X=2]$, and $\mathbb{E}[Y|X=3]$.

Given (C.11), evaluating (C.8) at $x = 0$ yields two equations that allow us to estimate μ_{00} and ρ_{0V}. This is because we already calibrated σ_0 and because $p(0,0)$ and $p(1,0)$ are observed in the data from the numbers of observations associated with each (Z, D, X) combination. Solving for these parameters yields $\mu_{00} = 0.0002505$ and $\rho_{0V} = -0.006081128$. We then arbitrarily set $\mu_{01} = 0.001$ and $\mu_{02} = -0.00075$ so that $m_0(\cdot)$ becomes concave in x.

Similarly, given (C.11), by equation (C.7) we have three equations. Since we already calibrated σ_1, we have four unknown parameters associated with $m_1(\cdot)$. We set $\rho_{1V} = 0.3$ to induce endogeneity in treatment because we obtained a very low value for ρ_{0V}, which enables us to use these three equations to solve for μ_1. We obtain $\mu_{10} = 0.002443738$, $\mu_{11} = 0.007943073$, and $\mu_{12} = -0.001791018$. Notice that these values imply that $m_1(\cdot)$ is also concave in x.

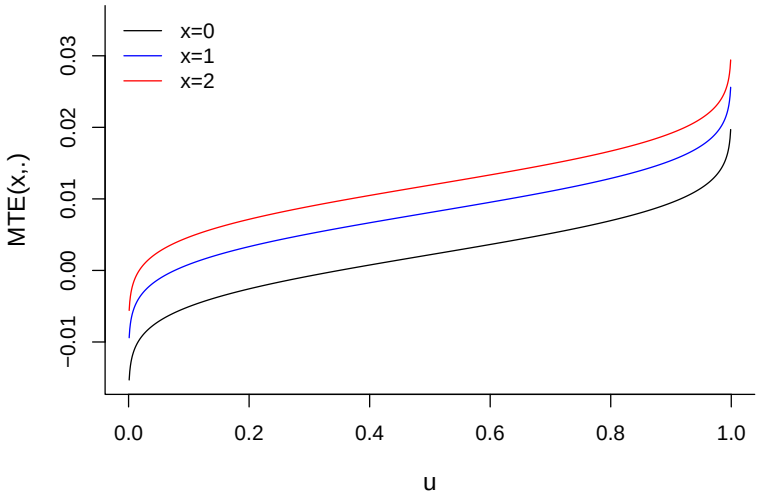
The resulting MTE function at $x \in \{0, 1, 2\}$ is displayed in Figure C.2. Because of additive separability, the MTE function does not cross at different values of x. While it is increasing in x, we can see that it is also concave.

C.5.4 Results

We now present the results of our calibrated simulations. We separate these results into two groups. First, we show the results from the sequential approach we described above. Second, we present the results from the procedure we described in Section 4, which ignores multiple ad exposures. We refer to this as the "naive approach." All results are shown in Figure C.3.

[5] We aggregated the quantities for $X \geq 3$ into a single quantity.

Figure C.2: MTE function with multiple exposures



Note: This figure shows the MTE function implied by the MTR functions from equation (C.11). Because we consider $S = 3$ impression opportunities and because x is the number of exposures prior to an impression opportunity, the MTE functions takes only three values in x.

Sequential approach

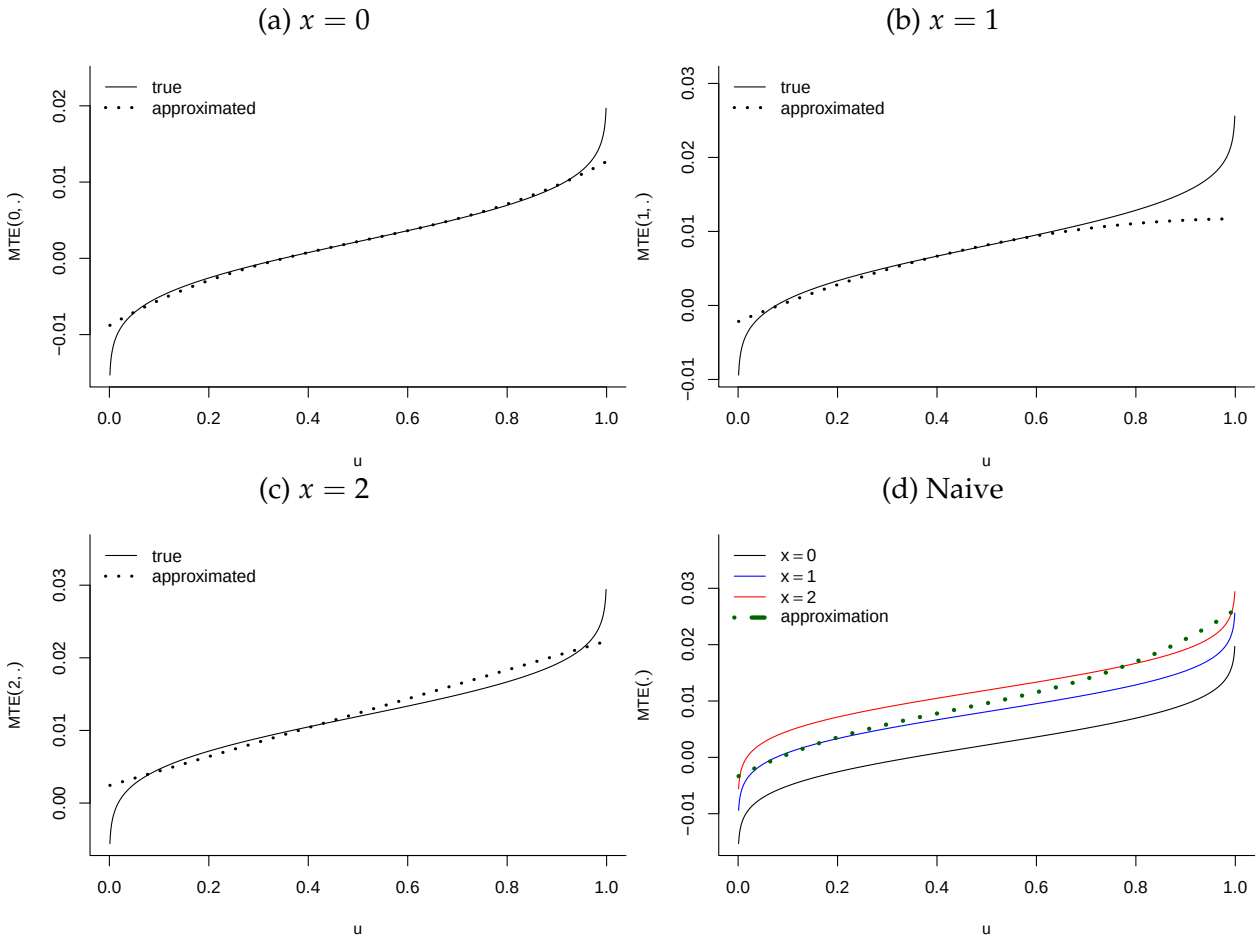
Figures C.3a−C.3c show the approximations to the MTE function at different values of x following the sequential procedure described above, depicting the MTE function with a solid line and the approximating polynomial with a dotted line. The method is able to approximate the MTE function well across all values of x even though, as discussed above, the degree of the approximating polynomial decreases as x increases.

Naive approach

Figure C.3d shows the MTE functions at different values of x in solid lines and the naive approximation with a dotted line. This naive approximation, which ignores multiple exposures, resembles an approximation to an "averaged" over x MTE function. Because the true function does not vary significantly between $x = 1$ and $x = 2$, this approximation is close to the true MTE function at these values, and especially at $x = 1$. However, the approximation to the MTE function at $x = 0$ is arguably poor.

The overall quality of the naive approximation is a reflection of the extent to which multiple exposures affect outcomes. Given our assumed DGP and calibration approach, repeated exposures do not seem to have a sizable impact. Since the outcome variable is an indicator for a purchase, this result is in accordance with results of the extant literature, which we discussed in Section 5.2. However, it is important to note that it is possible to choose a different DGP, also consistent with the observed data, where repeated exposures would have a strong impact on the outcome variable.

Figure C.3: Approximations to MTE function with multiple exposures



Note: Figures C.3a−C.3c plot the true MTE function for each value x with a solid black line and its sequential approximation with a dotted line. Figure C.3d replicates Figure C.2 and adds the naive approximation with a dotted line.

D Direct application of Brinch et al. (2017) with one cell

The method in Brinch et al. (2017) ("BMW") *can* be applied to data obtained from the typical experimental design. However, because this design yields one-sided noncompliance, $p(0) = 0$ while $p(1) \in (0, 1)$. Thus, equations (14) and (15) imply that the only moments identified from the data it provides are ψ_{11}, ψ_{01}, and ψ_{00}, where we omit the subscript c to ease notation since there is only one cell. Based on the logic from equations (16) and (17), these moments allow us to approximate $m_1(\cdot)$ with a constant function and $m_0(\cdot)$ with a linear function. Consequently the MTE function itself can be approximated with a linear function.

The ability to approximate the MTE function with a linear function might seem attractive, especially because it is not uncommon to maintain the assumption that the MTE function is indeed linear.[6] Nevertheless, the missingness of ψ_{10} implies that this approximation inherently features a restriction that nontrivially impacts not only the quality of the approximation, but also the structure of the endogeneity of treatment.

To see this, suppose for simplicity that $m_d(u) = \lambda_{d0} + \lambda_{d1}u$ for $d \in \{0, 1\}$. It follows that:

$$\begin{aligned}
\mathbb{E}\left[Y|D = 1, Z = z\right] &= \lambda_{10} + \frac{\lambda_{11}}{2}p(z) \\
\mathbb{E}\left[Y|D = 0, Z = z\right] &= \left(\lambda_{00} + \frac{\lambda_{01}}{2}\right) + \frac{\lambda_{01}}{2}p(z).
\end{aligned}$$

The data obtained from the typical experimental design thus enable us to recover λ_{00} and λ_{01}. Nevertheless, they do not allow us to recover λ_{10} and λ_{11} separately because we do not observe ψ_{10}. This is an underidentification problem: there are four parameters to be estimated (λ_{10}, λ_{11}, λ_{00}, and λ_{01}) but only three moments available to estimate such parameters (ψ_{11}, ψ_{01}, and ψ_{00}). To make progress, therefore, the researcher must impose an additional restriction on the parameters.

Unfortunately, we are unaware of any general theory or methodology that can provide clear guidance on such restriction, although progress has been recently made in this direction (e.g., Kowalski, 2023a). Whichever restriction is chosen, a key point is that such additional constraint must *always* be imposed to implement BMW's estimator using data collected from an experiment with one-sided noncompliance, which, to our knowledge, has not been noted in the literature. Given the pervasiveness of such experimental designs, we hope our approach provides a valuable solution to recover the MTE function more credibly.

[6] Examples of studies that maintained this linearity assumption are Olsen (1980), Moffitt (2008), French and Song (2014), Brinch et al. (2017), and Kowalski (2023b).

For the purposes of comparing our approach to a direct application of BMW, we consider four possible restrictions. First, in accordance with the approximation method from Section 3.3, we consider approximating $m_1(\cdot)$ with a constant, that is, imposing that $\lambda_{11} = 0$, which enables the estimation of λ_{10}. From a purely mechanical perspective, the higher $|\lambda_{11}|$ is, the lower the quality of the approximation. However, the constraint $\lambda_{11} = 0$ also has a deeper structural implication. It implies that all endogeneity stems from Y_0. This rules out certain forms of self-selection, such as the classic case of $D = \mathbb{1}\{Y_1 \geq Y_0\}$, where only units that benefit from treatment are treated. In our setting of online advertising, this precludes "perfect" ad exposure where units are exposed to the ad only if this benefits the advertiser.

The second restriction we consider it to set $\lambda_{11} = \lambda_{01}$, thereby ruling out endogeneity altogether but also allowing us to recover λ_{10}. However, this is often considered implausible, including in the context of online advertising.

Third, we impose $\lambda_{10} = 0$, which allows us to recover λ_{11}. Given linearity, this restriction implies that $m_1(\cdot)$ is either always negative or always positive, which can be justified in cases where Y_1 is bounded either from above or below at 0. The same can be achieved for the MTE function by imposing that $\lambda_{10} = \lambda_{00}$, while also enabling the estimation of λ_{11}.

Table D.1 also shows the optimal decisions and expected profit losses from using BMW's method directly under each of these constraints. We obtain the decisions using the procedure described in Section 3.4. We consider a sample size of 25,553,093, which corresponds to that of Study 4, use uniform priors, and take 1,000 draws to estimate the posterior means from equation (19).

The results do not show a systematic pattern. Notice that setting $\lambda_{10} = 0$ eliminates expected profit losses under DGP 1, which the multi-cell approach also does.

However, the decisions implied by the different versions of BMW do not vary across DGPs. To the extent that DGPs differ, and thus their implied optimal decisions, so will the performance of these different versions of BMW depending on what the true DGP is. As Figure 5 in the paper shows, the optimal solutions under DGPs 1 and 2 are quite different. Imposing $\lambda_{10} = 0$ would yield high losses under DGP 2, as expected. Under such DGP, the best version of the direct application of BMW among the ones we consider is the one that sets $\lambda_{10} = \lambda_{00}$; however, this version, unsurprisingly, performs poorly under DGP 1.

Table D.1: Expected profit losses (%) from different direct applications of BMW and from the multi-cell approach

DGP	True ϕ^*	Multi-cell		$\lambda_{11}=0$		$\lambda_{11}=\lambda_{01}$		$\lambda_{10}=0$		$\lambda_{10}=\lambda_{00}$	
		ϕ^*	Loss	ϕ^*	Loss	ϕ^*	Loss	ϕ^*	Loss	ϕ^*	Loss
1	1	1	0	0.484	72.30	0.433	79.04	1	0	0.661	45.38
2	0.755	0.762	0.03	0.484	39.10	0.433	50.70	1	58.00	0.662	5.92

Note: The optimal exposure rate, ϕ^*, is obtained using the procedure described in Section 3.4. We use a sample size of 25,553,093, which corresponds to that of Study 4, uniform priors, and 1,000 MCMC draws.

E Posterior distribution of approximation parameters

We now describe how to obtain draws from $f(\lambda_1, \lambda_0|\text{data})$. To this end, we need to set priors over ψ and p, denoted by $q(\psi, p)$, and the likelihood function of the data, $\ell(Y, D|\mathcal{C}, Z; \psi, p)$. We condition on Z and $\mathcal{C}$ because they are randomly chosen.

We need to consider two cases. First, notice that because $D = 0$ when $Z_c = 0$ for all c, we can pool all observations such that $Z_c = 0$ from all cells and use them to obtain the posterior distribution of ψ_{00}, as given in equation (12), conditional on the data. More precisely, given a prior distribution $q(\psi_{00})$ and the likelihood $\ell(Y|Z=0;\psi_{00})$, we can derive the posterior $f(\psi_{00}|Y, Z=0)$. The form of this distribution clearly depends on what type of variable Y is. In our application, Y is binary. Hence, for convenience we set:

$$\begin{aligned} Y|Z=0;\psi_{00} &\sim \text{Bernoulli}(\psi_{00}) \\ \psi_{00} &\sim \text{Beta}(\alpha_0, \beta_0) \end{aligned} \tag{E.1}$$

which implies that

$$\psi_{00}|Y, Z=0 \sim \text{Beta}\left(\alpha_0 + \sum_{i:Z_{ic}=0} Y_i, n_0 - \sum_{i:Z_{ic}=0} Y_i + \beta_0\right), \tag{E.2}$$

where n_0 is the total number of ineligible users.

The second case conditions on $\mathcal{C} = c$ and $Z_c = 1$. Denote the number of observations in this set by n_c. For convenience, we proceed in two steps, relying on the factorization:

$$\begin{aligned} \ell(Y, D|\mathcal{C}=c, Z_c=1; \psi_{11c}, \psi_{01c}, p(z_c)) = \ell(Y|D, \mathcal{C}=c, Z_c=1; \psi_{d1c}) \\ \times \ell(D|\mathcal{C}=c, Z_c=1; p(z_c)). \end{aligned} \tag{E.3}$$

First, recall that D is binary, so we proceed as above:

$$\begin{aligned}
&D|Z_c = 1, \mathcal{C} = c; p(z_c) \sim \text{Bernoulli}\left(p(z_c)\right) \\
&p(z_c) \sim \text{Beta}(\alpha_{Dc}, \beta_{Dc}) \\
&p(z_c)|D, Z_c = 1, \mathcal{C} = c \sim \text{Beta}\left(\alpha_{Dc} + \sum_{i:Z_{ic}=1, \mathcal{C}_i=c} D_i, n_c - \sum_{i:Z_{ic}=1, \mathcal{C}_i=c} D_i + \beta_{Dc}\right).
\end{aligned} \tag{E.4}$$

The second step consists of obtaining the posterior distribution of ψ_{d1c}, as given in equations (10) and (11), conditional on the data. Once again, the form of this distribution clearly depends on what type of variable Y is. Given our application, we proceed as in (E.1) and (E.2). Denote the number of observations such that $D = d$, $Z_c = 1$ and $\mathcal{C} = c$ by n_{d1c}. Then:

$$\begin{aligned}
&Y|D = d, Z_c = 1, \mathcal{C} = c; \psi_{d1c} \sim \text{Bernoulli}\left(\psi_{d1c}\right) \\
&\psi_{d1c} \sim \text{Beta}(\alpha_{d1c}, \beta_{d1c}) \\
&\psi_{d1c}|Y, D = d, Z_c = 1, \mathcal{C} = c \sim \text{Beta}\left(\alpha_{d1c} + \sum_{i:D_i=d, Z_{ic}=1, \mathcal{C}_i=c} Y_i, n_{d1c} - \sum_{i:D_i=d, Z_{ic}=1, \mathcal{C}_i=c} Y_i + \beta_{d1c}\right).
\end{aligned} \tag{E.5}$$

The approach we take is sequential: first, we obtains draws of p and then, conditional on them, we draw ψ. Consequently, we avoid the feedback issue studied in Zigler et al. (2013).

F Distributing experiment budget across multiple cells

We formally describe how the experimenter can distribute the experiment budget across multiple cells to generate the variation necessary to estimate the parameters of interest.

First, consider a two-cell experiment, and fix $\Pr(\mathcal{C} = 1)$ and $\Pr(Z_1 = 1|\mathcal{C} = 1)$. Let τ_1 be the fraction of the original budget, B, allocated to Cell 1, so that $B_1 = \tau_1 \times B$. The effective budget for Cell 1 is then $\frac{B_1}{\Pr(Z_1=1|\mathcal{C}=1)\times\Pr(\mathcal{C}=1)} = \frac{\tau_1}{\Pr(Z_1=1|\mathcal{C}=1)\times\Pr(\mathcal{C}=1)} \times B$, so that $p(Z_1 = 1) = \kappa^{-1}\left(\frac{\tau_1}{\Pr(Z_1=1|\mathcal{C}=1)\times\Pr(\mathcal{C}=1)} \times B\right)$. Equivalently, we then obtain $B_2 = (1 - \tau_1) \times B$, so that the effective budget for Cell 2 becomes $\frac{B_2}{\Pr(Z_2=1|\mathcal{C}=2)\times(1-\Pr(\mathcal{C}=1))} = \frac{1-\tau_1}{\Pr(Z_2=1|\mathcal{C}=2)\times(1-\Pr(\mathcal{C}=1))} \times B$ and $p(Z_2 = 1) = \kappa^{-1}\left(\frac{1-\tau_1}{\Pr(Z_2=1|\mathcal{C}=2)\times(1-\Pr(\mathcal{C}=1))} \times B\right)$.

Hence, when running a two-cell design, given B and $\kappa(\cdot)$, the experimenter has four de-

cision variables: $\Pr(\mathcal{C}=1)$, τ_1, $\Pr(Z_1=1|\mathcal{C}=1)$, and $\Pr(Z_2=1|\mathcal{C}=2)$. For the desired variation in the propensity score to be generated, it is necessary that $\frac{\tau_1}{\Pr(Z_1=1|\mathcal{C}=1)\times\Pr(\mathcal{C}=1)} \neq \frac{1-\tau_1}{\Pr(Z_2=1|\mathcal{C}=2)\times(1-\Pr(\mathcal{C}=1))}$. Thus, the experimenter can always guarantee that this constraint is satisfied.

This framework can be generalized to C cells in a straightforward manner. In this case, the experimenter has $3C-2$ variables: $\{\tau_c, \Pr(\mathcal{C}=c)\}_{c=1}^{C-1}$ and $\{\Pr(Z_c=1|\mathcal{C}=c)\}_{c=1}^{C}$, under the constraints that, for all $c=1,\cdots,C$, $\tau_c \in [0,1]$, $\Pr(Z_c=1|\mathcal{C}=c) \in [0,1]$ and $\Pr(\mathcal{C}=c) \in [0,1]$, plus $\sum_{c=1}^{C}\tau_c = 1$ and $\sum_{c=1}^{C}\Pr(\mathcal{C}=c)=1$. To ensure that the propensity scores differ from one another, it is then required that $\frac{\tau_c}{\Pr(Z_c=1|\mathcal{C}=c)\times\Pr(\mathcal{C}=c)} \neq \frac{\tau_{c'}}{\Pr(Z_{c'}=1|\mathcal{C}=c')\times\Pr(\mathcal{C}=c')}$ for all $c \neq c'$.

G A more complex DGP

We consider a more complex MTE function to dig deeper into the ability of our proposed multi-cell experimental design to provide a good approximation of this function and to inform decision-making. In particular, we assess how the performance of our approach changes as the number of cells increases.

G.1 New MTE and expected profit functions

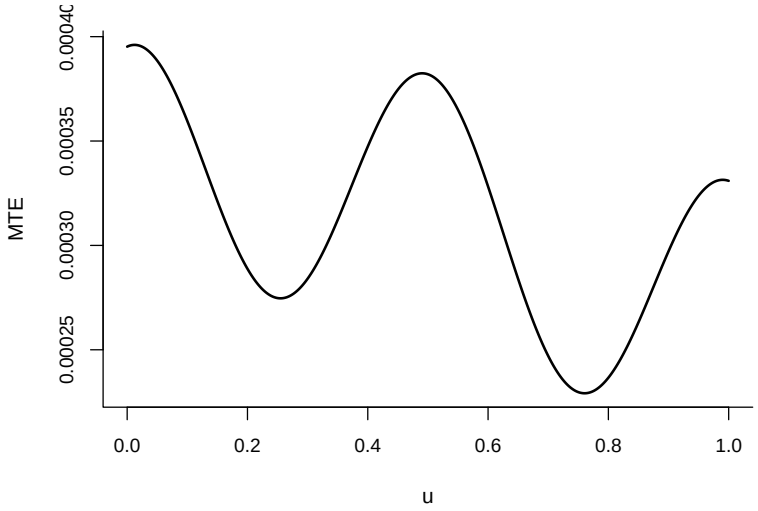
We choose $m_1(u) = \beth\frac{1}{1+u}$ and $m_0(u) = \gimel\frac{1}{(1+u)^2} + \daleth\sin^2(2\pi u)$. As before, the parameters $\beth$, $\daleth$, and $\gimel$ are computed to match the observed ψ_{11}, ψ_{01}, and ψ_{00}. Figure G.1 depicts the resulting MTE function. We consider this function to be "complex" because is not monotonic, concave, or convex over the entire domain.

Figure G.2 shows the implied expected profit function under our original cost function: $\kappa(\phi) = 0.001\phi^4$. Under this DGP, the optimal decision is to treat 45.5% of the population.

G.2 Approximations based on different numbers of cells

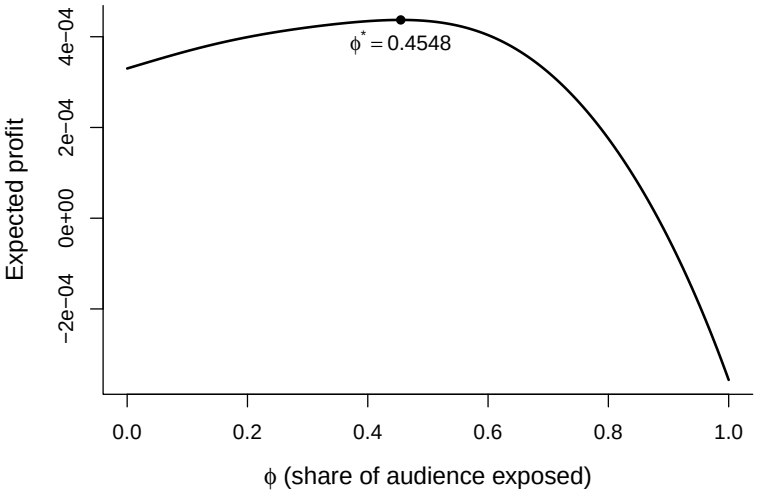
We now study how the quality of approximation changes as more cells are included into the experiment. In particular, in addition to the two-cell design, we consider three-cell and five-cell designs, which allow us to obtain cubic and quintic approximations to the MTE function, respectively. Cells 1 and 2 display the same eligibility probabilities and propensity scores as in Section 4. Cell 3 has $\Pr(Z_3=1|\mathcal{C}=3)=0.5$ and $p(Z_3=1)$,

Figure G.1: $\text{MTE}(u) = \gimel \frac{1}{1+u} - \beth \frac{1}{(1+u)^2} - \daleth \sin^2(2\pi u)$



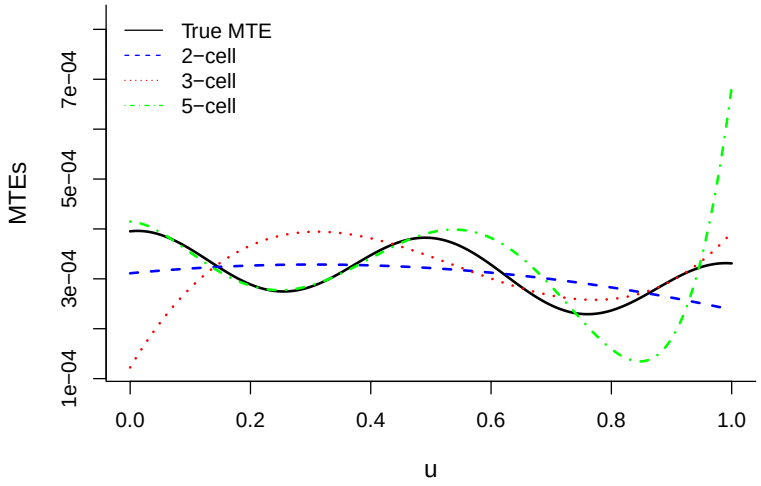
Note:

Figure G.2: Expected profit function under complex MTE



Note: This figure shows the expected profit function implied by the MTE function shown in Figure G.1 combined with the cost function $\kappa(\phi) = 0.001\phi^4$.

Figure G.3: Approximations to complex MTE function



Note: This figure shows approximations to the MTE function shown in Figure G.1 for different numbers of cells. The solid black line is the true MTE function, the dashed blue line is the two-cell approximation, the dotted red line is the three-cell approximation, and the dashed-dotted green line is the five-cell approximation.

chosen so that $\Pr(Z_1 = 1|\mathcal{C} = 1) \times p(Z_1 = 1) = \Pr(Z_3 = 1|\mathcal{C} = 3) \times p(Z_3 = 1)$. We set $\Pr(Z_4 = 1|\mathcal{C} = 4) = 0.25$ and $\Pr(Z_5 = 1||\mathcal{C} = 5) = 0.9$, and set $p(Z_5 = 1)$ so that $\Pr(Z_1 = 1|\mathcal{C} = 1) \times p(Z_1 = 1) = \Pr(Z_5 = 1|\mathcal{C} = 5) \times p(Z_5 = 1)$, obtaining $p(Z_5 = 1) \approx 0.288$. Finally, we pick $p(Z_4 = 1) = 0.17$ to increase the range of values covered by the propensity scores. Figure G.3 shows the four approximated MTE functions along with the true one.

Seemingly, as the number of cells increases, so does the quality of the approximation. Nevertheless, we note that at the extreme points of the domain these approximations might worsen because of the extrapolation imposed by the polynomial functional form. To quantify and assess its impact, we compute the distance between the true and approximated MTE functions using the sup-norm, L_2-norm, and the Relative ATE error. These quantities are shown in Table G.1, along with the optimal decisions implied by these approximations and the losses in expected profit they imply.

The different criteria demonstrate that, in this example, no single approximation strictly dominates the others. Interestingly, according to the criteria that measure discrepancy between the entire MTE function and its approximation, the simplest approximation, with only two cells, performs best. In turn, the five-cell approximation is the one whose implied ATE is closest to the true ATE and that implies the smallest loss in expected profit. Caution is required if wishing to generalize these results because we do not know whether the assumed MTE functional form bears any resemblance to one that might be encountered in the real-world.

Table G.1: Closeness to MTE function and expected profit losses (%) as a function of the number of cells

Method	sup-norm	L_2-norm	Relative ATE error	ϕ^*	Loss
Two-cell	0.00009	0.00005	-0.02445	0.434	0.10
Three-cell	0.00027	0.00008	0.00585	0.545	2.60
Five-cell	0.00035	0.00006	0.00144	0.696	25.51
True	—	—	—	0.455	—

Note: This table shows the three distance metrics between the true and the approximated MTE functions given in equation (20). The optimal exposure rate, ϕ^*, is obtained using the procedure described in Section 3.4. We use a sample size of 25,553,093, which corresponds to that of Study 4, uniform priors, and 1,000 draws to estimate the posterior means from equation (19).

These results further highlight that, even though they are connected, the tasks of approximating the MTE function and approximating the maximum of the expected profit function are not perfectly aligned. The extent to which these approximations are misaligned depends on the underlying MTE and cost functions.

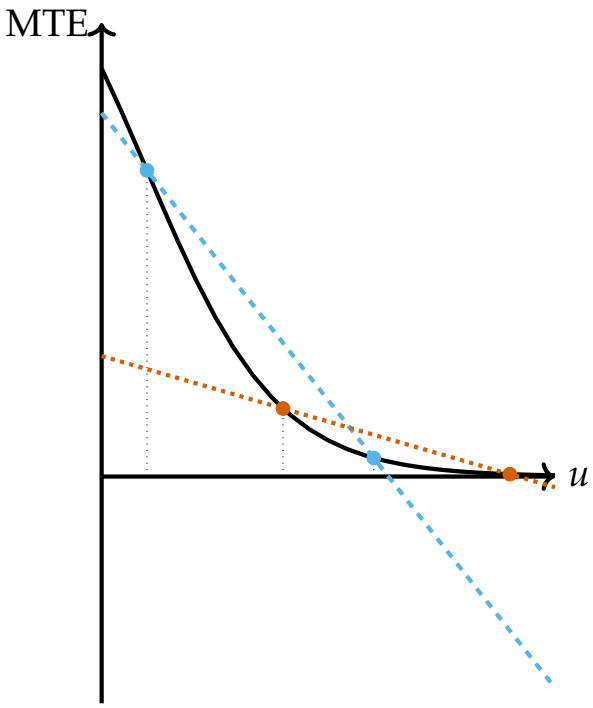
H MTE function under different assumptions

We consider two commonly made assumptions on the DGP to assess whether they impose enough structure on the resulting MTE function to imply specific guidance on how to choose the number of cells or values for the propensity score during the experiment. Under each assumption, we present an example such that no specific guidance is obtained.

H.1 Monotonic MTE function

One possible assumption the researcher might be willing to make is that the MTE function is monotonic. This is necessarily satisfied when the MTE function is linear or when the joint distribution of potential outcomes and the selection unobservable is jointly, which are commonly made assumptions. We now provide an example where monotonicity of the MTE function does not necessarily aid in choosing C or $p(Z_c = 1)$.

Figure H.1: Different linear approximations for an MTE curve



To this end, assume that $\text{MTE}(u) = \frac{2.7}{2+2^{10u}}$. This function is not only monotonic, but it also is strictly convex and nonnegative. Hence, it is a fairly "well-behaved" function.

In particular, the monotonicity might suggest that a linear approximation might be satisfactory. However, H.1 shows that not to be the case because a linear approximation might cover a wide range of negative values, which is a marked difference from the true MTE function. Furthermore, a linear approximation is highly susceptible to values of the propensity score. As Figure H.1 shows, depending on these values the resulting linear curve can have very distinct slopes.

Given this MTE function, it might seem like having one additional cell might suffice to obtain a satisfactory approximation. Nevertheless, this need not be the case. Figure H.2 depicts two quadratic approximations to the MTE function. As we can see, neither approximation is particularly good, and one fails to capture the monotonicity of the MTE function. This is a result of the range of values taken by the propensity score, which, in both cases, is limited.

On the other hand, as perhaps expected, when the propensity score covers a wider range of values, the quality of the approximation can be higher, as we show in Figure H.3. This results echoes the ones from our analysis in Section 4.4. Note, however, that this approximation, unlike the true MTE function, displays negative values.

Figure H.2: Different quadratic approximations for an MTE curve

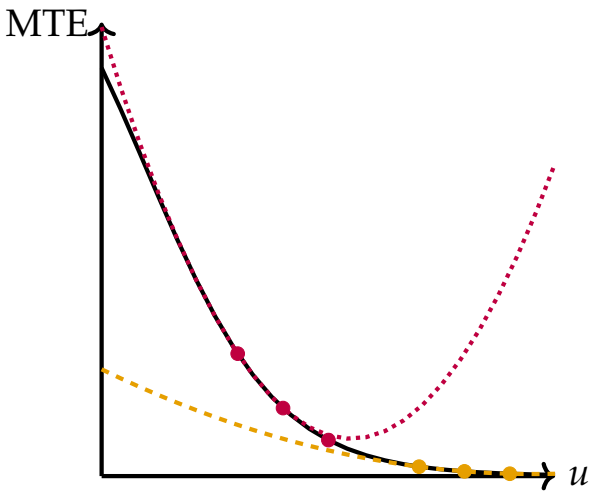


Figure H.3: Different quadratic approximations for an MTE curve

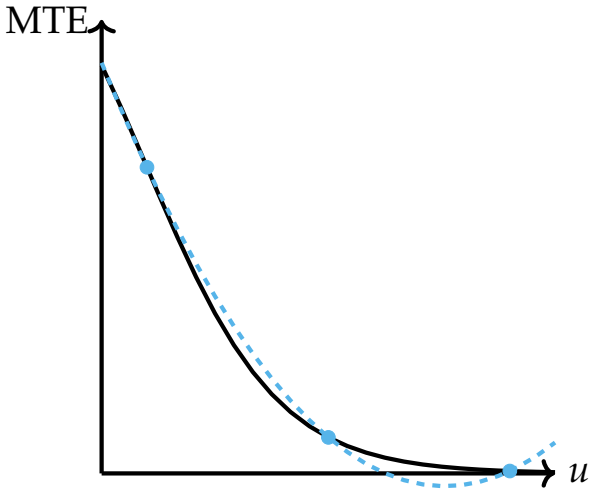
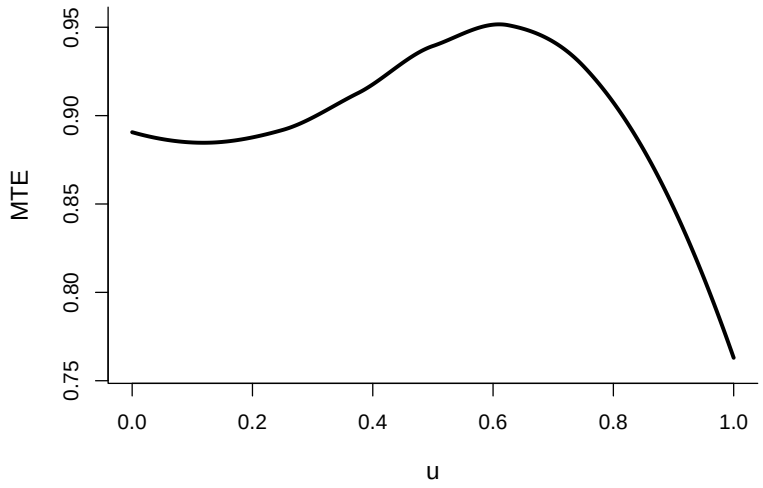


Figure H.4: MTE function under monotone treatment response



A priori, it is unclear, however, the curvature of the function, and therefore what the exact range of propensity score values should be.

H.2 Monotone treatment response

A different assumption is that of monotone treatment response (Manski, 1997). It implies that the treatment effects themselves always have the same sign, which implies that so does the MTE function. However, the previous example suggests that this will not suffice to make this MTE function sufficiently "well-behaved" for us to obtain precise guidance for the specific design of the experiment. Indeed, the example below confirms this.

Assume that $Y_0|U = u \sim N\left(0.9u - 3.8u^2 + 3.3u^3, 1\right)$ and $Y_1|Y_0 \sim TN\left(0, 1, Y_0, +\infty\right)$, so that Y_1 follows a standard normal distribution truncated from below at Y_0. The resulting MTE function is shown in Figure H.4. Unsurprisingly, monotone treatment response ensures that the function always has the same sign, but is not even enough to impose, for instance, monotonicty. Consequently, on its own this assumption is not helpful in informing specifically how the experiment should be designed and implemented.

References

Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond LATE with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039.

French, E. and Song, J. (2014). The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy*, 6(2):291–337.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., and Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, 38(2):193–225.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics*, 88(3):389–432.

Heckman, J. J. and Vytlacil, E. (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics, Vol. 6B*, pages 4875–5143. Elsevier.

Kowalski, A. E. (2023a). Behavior within a clinical trial and implications for mammography guidelines. *Review of Economic Studies*, 90(1):432–462.

Kowalski, A. E. (2023b). Reconciling seemingly contradictory results from the Oregon Health Insurance Experiment and the Massachussets Health Reform. *Review of Economics and Statistics*, 105(3):646–664.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, 65(6):1311–1334.

Moffitt, R. (2008). Estimating marginal treatment effects in heterogeneous populations. *Annales d'Économie et Statistique*, 91/92:239–261.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619.

Mogstad, M. and Torgovitsky, A. (2018). Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics*, 10:577–613.

Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica*, 48(7):1815–1820.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics*, 69(1):263–273.