Supplemental Appendix to Approaches to Statistical Efficiency when comparing the embedded adaptive interventions in a SMART

Appendices

A Common Primary Aims

Consider the following marginal structural mean model, i.e. , $\mu^{(a_1,a_{2NR})}(\boldsymbol{\gamma})$ for a prototypical SMART as follows:

$$\mu^{(a_1,a_{2NR})}(\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 a_1 + \gamma_2 a_{2NR} + \gamma_3 a_1 a_{2NR},$$

where $a_1, a_{2NR} \in \{-1, 1\}$. This corresponds to the marginal structural mean model found in Equation 1 in the manuscript.

Table 1 shows each of the most common primary aims in a prototypical SMART. The first is the main effect for the first-stage intervention. The second common primary aim is the main effect of the second-stage intervention among the non-responders. The final six are comparisons of embedded adaptive interventions. For each of these primary aims, Table 1 provides the causal effect along with a regression formulation for that causal effect using the marginal structural mean model found above.

Primary Aim	**Causal Effect**	**Regression Formulation**
Main effect of 1st Stage	$\mathbb{E}[Y(1, A_{2NR}) - Y(-1, A_{2NR})]$	$2\gamma_1$
Main effect of 2nd Stage among NR	$\mathbb{E}[Y(A_1, 1) - Y(A_1, -1) \mid R(A_1) = 0)]$	$2\gamma_2 \times (1 - R)^{-1}$
AI Comparison: $(1, 1)$ vs. $(1, -1)$	$\mathbb{E}[Y(1, 1) - Y(1, -1)]$	$2(\gamma_2 + \gamma_3)$
AI Comparison: $(1, 1)$ vs. $(-1, 1)$	$\mathbb{E}[Y(1, 1) - Y(-1, 1)]$	$2(\gamma_1 + \gamma_3)$
AI Comparison: $(1, 1)$ vs. $(-1, -1)$	$\mathbb{E}[Y(1, 1) - Y(-1, -1)]$	$2(\gamma_1 + \gamma_2)$
AI Comparison: $(1, -1)$ vs. $(-1, 1)$	$\mathbb{E}[Y(1, -1) - Y(-1, 1)]$	$2(\gamma_1 - \gamma_2)$
AI Comparison: $(1, -1)$ vs. $(-1, -1)$	$\mathbb{E}[Y(1, -1) - Y(-1, -1)]$	$2(\gamma_1 - \gamma_3)$
AI Comparison: $(-1, 1)$ vs. $(-1, -1)$	$\mathbb{E}[Y(-1, 1) - Y(-1, -1)]$	$2(\gamma_2 - \gamma_3)$

Table 1: Each of the primary aims for a prototypical SMART along with their causal effects and the corresponding regression-based formulation for these causal effects using the marginal structural mean model found above.

B Derivations of a saturated marginal mean model

Take the following saturated model:

$$\mu^{(a_1,a_{2NR})} = \mathbb{E}[Y^{(a_1,a_{2NR})}] = \mathbb{1}^{(a_1=1,a_{2NR}=1)}\mu^{(1,1)} + \mathbb{1}^{(1,-1)}\mu^{(1,-1)} + \mathbb{1}^{(-1,1)}\mu^{(-1,1)} + \mathbb{1}^{(-1,-1)}\mu^{(-1,-1)}.$$

We can estimate $\hat{\mu}$, the (4×1) vector of parameters, by solving the following estimating equations:

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}\sum_{d}[\sigma^{-2}(d)\mathbb{1}^{(d)}(A_{1,i}, R_i, A_{2,i})W^{(d)}(A_{1,i}, R_i, A_{2,i})D^{(d)T}(Y_i - \mu^{(d)}(\hat{\mu})) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}_i(A_{1,i}, R_i, A_{2,i}, Y_i; \hat{\mu}), \end{aligned} \tag{1}$$

where $D^{(d)}$ denotes the Jacobian of $\mu^{(d)}(\mu)$ with respect to μ and the scalar parameter σ^2 is a working model for the variance of the potential outcome $Y^{(d)}$. Importantly, this is a working assumption; it is not necessary for $\sigma^2 = \mathrm{Var}(Y^{(d)})$ to hold for the derivations shown here. For the time being, we assume that the variance is homogeneous across the various adaptive interventions and thus write $\sigma(d)^2 = \sigma^2$

We now perform a first-order Taylor expansion:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}(\cdot) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{M}(\cdot)}{\partial \mu}(\hat{\mu} - \mu_0) + o_p(1).$$

We rearrange, leading to the following:

$$\sqrt{n}(\hat{\mu} - \mu_0) = -\mathbb{E}\Big[\frac{\partial \mathbf{M}(\cdot)}{\partial \mu}\Big]^{-1}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{M}_i\Big) + o_p(1).$$

Now let us denote $\mathbb{E}\Big[\frac{\partial \mathbf{M}(\cdot)}{\partial \mu}\Big]$ as follows:

$$\mathbb{E}\Big[\frac{\partial \mathbf{M}(\cdot)}{\partial \mu}\Big] = \frac{1}{n}\sum_{i=1}^{n}\sum_{d}[\sigma^{-2}\mathbb{1}^{(d)}(A_{1,i}, R_i, A_{2,i})W^{(d)}(A_{1,i}, R_i, A_{2,i})D^{(d)}D^{(d)T}]$$
$$= \mathbf{B}(\cdot).$$

Then the variance can be written as follows: $\text{Var}\left(\sqrt{n}(\hat{\mu} - \mu_0)\right) = \mathbf{B}^{-1}\mathbb{E}[\mathbf{M}\mathbf{M}^T]\mathbf{B}^{-1}$.

Now let us decompose $\mathbf{B}$ (the bread) and $\mathbf{M}\mathbf{M}^T$ (the meat). We begin with the bread. $D^{(d)}$ is a vector of indicator functions denoting consistency with AI d. Thus, $\mathbf{B}$ is simply a $(4{\times}4)$ matrix where each entry (j, k) is a weighted sum of the number of individuals consistent with AI j and AI k, divided by an estimate of the variance σ^2: $\sum_{i=1}^{n} W_i \mathbb{1}_i^{(k)}\mathbb{1}_i^{(l)}/\sigma^2$.

Now let us examine $\mathbf{M}\mathbf{M}^T$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{d}\sigma^{-2}\mathbb{1}_i^{(d)}W_i D^{(d)T}(Y_i - \mu^{(d)}(\mu))(Y_i - \mu^{(d)}(\mu))^T D^{(d)}W_i\mathbb{1}_i^{(d)}\sigma^{-2}.$$

We start by analyzing the interior: $(Y_i - \mu^{(d)}(\hat{\mu}))$ is a vector of the differences between Y_i and the given mean $\mu^{(d)}$ $\forall$ d, i.e. $\left(Y_i - \hat{\mu}^{(1,1)}, Y_i - \hat{\mu}^{(1,-1)}, Y_i - \hat{\mu}^{(-1,1)}, Y_i - \hat{\mu}^{(-1,-1)}\right)$. Thus, $(Y_i - \mu^{(d)}(\hat{\mu}))(Y_i - \mu^{(d)}(\hat{\mu}))^T$ is a (4×4) matrix of the cross product of the residuals under each AI: the (k, l) entry of this matrix is $(Y_i - \mu^{(k)})(Y_i - \mu^{(l)})$.

The indicators in the meat function set some entries to zero such that only terms consistent with a given AI contribute. Thus, the (k, l) entry of the meat is as follows: $\sum_{i=1}^{n} W_i^2\left(\mathbb{1}_i^{(k)}(Y_i - \hat{\mu}^{(k)})\right)\left(\mathbb{1}_i^{(l)}(Y_i - \hat{\mu}^{(l)})\right)$.

We then incorporate the bread matrix, leaving us with the following entries in our covariance matrix:

$$\text{Var}(\mu^{(k)}) = \frac{\sum_{i=1}^{n} W_i^2\mathbb{1}_i^{(k)})(Y_i - \mu^{(k)})^2}{\sum_{i=1}^{n} W_i^2\mathbb{1}_i^{(k)}},$$

the diagonal entries, and

$$\text{Cov}(\mu^{(k)}, \mu^{(l)}) = \frac{\sum_{i=1}^{n} W_i^2\mathbb{1}_i^{(k)}(Y_i - \mu^{(k)})\mathbb{1}_i^{(l)}(Y_i - \mu^{(l)})}{\sum_{i=1}^{n} W_i^2\mathbb{1}_i^{(k)}\mathbb{1}_i^{(l)}},$$

the off-diagonal entries. Note that the working model for the variance, σ^2, cancelled out.

C Repeated Measurements with Non-Independent Working Variance Structure

In this section, we will show that under an appropriate weighting structure, we can use a non-independent working variance structure despite repeated measurements without bias. We break this section up into two parts. In Section C.1, we show that the estimating equation described in our manuscript will be unbiased when applying a non-independent working variance structure so long as we use the specific form of weights proposed in our manuscrupt. In Section C.2 we show that the estimating equation described in our manuscript will be biased when applying a non-independent working variance structure under an alternative, yet common approach to weighting.

To simplify our argument, we remove time zero from our repeated measurements model, instead using a two-time point model. Thus, our repeated measures marginal structural model has the following saturated structure:

$$\mu = \mathbb{1}(t = 1)(\beta_1 + \beta_2 a_1) + \mathbb{1}(t = 2)(\beta_3 + \beta_4 a_1 + \beta_5 a_2 + \beta_6 a_1 a_2).$$

In addition, we use an unrestricted SMART (Nahum-Shani and Almirall, 2019) in order to reduce the algebraic burden involved. Nonetheless, the argument presented in this appendix is readily extendable to other SMART settings including full-factorial and prototypical SMARTs.

For the purpose of this example, we use binary coding (0/1) rather than contrast coding (-1/1) for both first and second-stage treatments. Let us first define our first and second stage weights as follows:

$$W_1 = W_1(A_1) = A_1 \frac{1}{\mathbb{P}(A_1 = 1)} + (1 - A_1) \frac{1}{1 - \mathbb{P}(A_1 = 1)},$$

$$W_2 = W_2(A_1, A_2) = A_2 \frac{1}{\mathbb{P}(A_2 = 1|A_1)} + (1 - A_2) \frac{1}{1 - \mathbb{P}(A_2 = 1|A_1)}.$$

For ease of notation, let us define $p_1 = \mathbb{P}(A_1 = 1)$ and $p_2 = \mathbb{P}(A_2 = 1|A_1)$. Note that we are using general forms for the weights to show that we are not limited to 50%/50% randomization probabilities. Under our weighting scheme, the weights we apply to observations at each time point are equal to the product of the weights, i.e., $W = \text{diag}(W_1W_2, W_1W_2)$. Following the approach suggested in Boruvka et al. (2018) and elsewhere, these weights would instead be $W = \text{diag}(W_1, W_1W_2)$.

Based on the marginal structural model μ, D^T is a 6×2 matrix where $d_{1,A_1} = (1, A_1, 0, 0, 0, 0)^T$ and $d_{2,\bar{A}_2} = (0, 0, 1, A_1, A_2, A_1A_2)^T$ and $(Y - D\beta) = (Y_1 - d_{1,A_1}\beta, Y_2 - d_{2,\bar{A}_2}\beta)^T$. We also define V to be an exchangeable (2×2) matrix with ρ in the off-diagonal.

For a two-time point repeated measurements model, our estimating equation can be defined as follows:

$$\begin{aligned} 0 = \mathbb{E}(&d_{1,A_1}W_1W_2(Y_1 - d_{1,A_1}\beta) + d_{2,\bar{A}_2}W_1W_2(Y_2 - d_{2,\bar{A}_2}\beta) - \\ &\rho\mathbb{E}(d_{1,A_1}W_1W_2(Y_2 - d_{2,\bar{A}_2}\beta) + d_{2,\bar{A}_2}W_1W_2(Y_1 - d_{1,A_1}\beta)). \end{aligned}$$

Note that when we use an independent working variance structure, $\rho = 0$ and the third and fourth terms (i.e., those terms in the second expectation) disappear. We know that the remainder (i.e., the first and second terms) is mean-zero and thus, this estimating equation is unbiased. In order for this estimating equation to be unbiased when $\rho \neq 0$, then the two terms within the second expectation must have mean zero as well.

C.1 Part 1

We begin with the first term in the second expectation, i.e., $\mathbb{E}[d_{1,A_1}W_1W_2(Y_2-d_{2,\bar{A}_2}\beta)]$. First, note that $d_{1,A_1}W_1 = A_1d_{1,1}/p_1+(1-A_1)d_{1,0}/(1-p_1)$. As a result, $\mathbb{E}[d_{1,A_1}W_1] = d_{1,1}+d_{1,0} = \kappa_1$. where κ_1 is a constant.

Now, $\mathbb{E}[d_{1,A_1}W_1W_2(Y_2 - d_{2,\bar{A}_2}\beta)] = \mathbb{E}_{A_1,A_2}[\mathbb{E}_{Y_2}(d_{1_{A_1}}W_1W_2(Y_2 - d_{2,\bar{a}_2}\beta)|A_1, A_2)]$. Note that from our assumptions about consistency and sequential ignorability, $\mathbb{E}(Y_2|A_1, A_2) = \mathbb{E}(Y_2(A_1, a_2)|A_1)$. Therefore, it then follows that $\mathbb{E}_{A_1,A_2}[\mathbb{E}_{Y_2}(d_{1,A_1}W_1W_2(Y_2-d_{2,\bar{a}_2}\beta)|A_1, A_2)] = \mathbb{E}_{A_1,A_2}[\mathbb{E}_{Y_2}(d_{1,A_1}W_1W_2(Y_2(A_1, a_2) - d_{2,\bar{a}_2}\beta)|A_1, A_2)]$. Using the law of iterated expectations, this equals $\mathbb{E}[W_1W_2(Y_2(A_1, a_2) - d_{2,\bar{a}_2})]$.

We again apply the law of iterated expectations, allowing us to re-write the RHS as $\mathbb{E}_{A_1}[\mathbb{E}_{Y_2(A_1,a_2),A_2}(d_{1,A_1}W_1W_2(Y_2 - d_{2,\bar{a}_2}\beta)|A_1)]$. Note that $\mathbb{E}[W_2|A_1] = p_2/p_2 + (1 - p_2)/(1 - p_2) = 2$ and thus, we can re-write this as $2\mathbb{E}_{A_1}[\mathbb{E}_{Y_2(A_1,a_2)}(d_{1,A_1}W_1(Y_2(A_1, a_2) - d_{2,\bar{a}_2}\beta)|A_1)]$. Now through consistency and sequential ignorability, $\mathbb{E}(Y_2(A_1, a_2)|A_1) = \mathbb{E}(Y_2(a_1, a_2))$, so the RHS now equals $2\mathbb{E}_{A_1}[\mathbb{E}_{Y_2(A_1,a_2),A_2}(d_{1,A_1}W_1(Y_2(a_1, a_2) - d_{2,\bar{a}_2}\beta))] = 2\mathbb{E}[d_{1,A_1}W_1(Y_2(a_1, a_2) - d_{2,\bar{a}_2}\beta)]$.

From above, $\mathbb{E}[d_{1,A_1}W_1] = \kappa_1$ and thus, this simplifies to $2\kappa_1\mathbb{E}_{Y_2(a_1,a_2)}[Y_2(a_1, a_2) - d_{2,\bar{a}_2}\beta]$, which has mean zero. Note that in restricted SMART settings like the prototypical SMART or the full-factorial design, we would have an additional set of analogous steps between when we average out A_2 and when we average out A_1 where we handle response status R.

We now examine the next term in the second expectation: $\mathbb{E}[d_{2,\bar{A}_2}W_1W_2(Y_1 - d_{1,A_1}\beta)]$. Through the law of iterated expectation, this equals $\mathbb{E}_{A_1}[\mathbb{E}_{Y_1,A_2}(d_{2,\bar{A}_2}W_1W_2(Y_1-d_{1,a_1}\beta)|A_1)]$. Now, note that $\mathbb{E}(d_{2,\bar{A}_2}W_2|A_1) = d_{2,a_1,1} + d_{2,a_1,0} = \kappa_2$, a constant. Thus, we can rewrite this as follows: $\mathbb{E}_{A_1}[\kappa_2\mathbb{E}_{Y_1}(W_1(Y_1 - d_{1,a_1}\beta)|A_1)]$.

In addition, note that through consistency and sequential ignorability, $\mathbb{E}_{Y_1}(Y_1|A_1) = \mathbb{E}_{Y_1(a_1)}(Y(a_1))$ so we further simplify: $\mathbb{E}_{A_1}[\kappa_2\mathbb{E}_{Y_1(a_1)}(W_1(Y_1(a_1)-d_{1,a_1}\beta)|A_1)]$. We now average over A_1 through the law of iterated expectations: $\kappa_2\mathbb{E}_{Y_1(a_1),A_1}[W_1(Y(a_1) - d_{1,a_1}\beta)]$.

Note that $\mathbb{E}(W_1) = p_1/p_1 + (1-p_1)/(1-p_1) = 2$ and thus, we simplify to $2\kappa_2 \mathbb{E}_{Y(a_1)}[Y(a_1) - d_{1,a_1}\beta]$, which has mean zero. Thus, when we apply our alternative weighting scheme, this estimating equation will be unbiased when when we use a non-independent working variance structure.

C.2 Part 2

Here, we will show that if we use the alternative weighting scheme used in Boruvka et al. (2018), i.e., we define $W = \text{diag}(W_1, W_1 W_2)$ rather than $W = \text{diag}(W_1 W_2, W_1 W_2)$, this final term will be biased.

For a two-time point repeated measurements model, our estimating equation can now be defined as follows:

$$
\begin{aligned}
0 = \mathbb{E}(d_{1,A_1} W_1 (Y_1 - d_{1,A_1}\beta) + d_{2,\bar{A}_2} W_1 W_2 (Y_2 - d_{2,\bar{A}_2}\beta) - \\
\rho \mathbb{E}(d_{1,A_1} W_1 W_2 (Y_2 - d_{2,\bar{A}_2}\beta) + d_{2,\bar{A}_2} W_1 (Y_1 - d_{1,A_1}\beta)).
\end{aligned}
$$

The first expectation is still mean-zero. Likewise, the first term in the second expectation, $\mathbb{E}[d_{1,A_1} W_1 W_2 (Y_2 - d_{2,\bar{A}_2}\beta)]$, is unchanged and thus mean-zero as shown in Section C.1.

Thus, bias must enter through the final term, $d_{2,\bar{A}_2} W_1 (Y_1 - d_{1,A_1}\beta)$, when we use the alternative weighting scheme. We see this as follows: $\mathbb{E}[d_{2,\bar{A}_2} W_1 (Y_1 - d_{1,A_1}\beta)] = \mathbb{E}_{A_1}[\mathbb{E}_{Y_1,A_2}(d_{2,\bar{A}_2} W_1 (Y_1 - d_{1,A_1}\beta)|A_1)]$. But note that $\mathbb{E}_{Y_1,A_2}[d_{2,\bar{A}_2} W_1 | A_1] = 2(d_{2,a_1,1} p_2 + d_{2,a_1,0}(1-p_2))$, which remains a function of A_1 through p_2, even after conditioning on A_1. Thus, it is not possible to ensure this term is mean zero.

D Variance-Covariance Matrix for ASIC Simulations

This is the variance-covariance matrix for our baseline covariates. The covariates are:

- Number of students >500.

- Percentage of students on free/reduced price lunch.
- Rural versus urban school.
- Total school professional level of education.
- Total school professional tenure.
- Delivered pre-randomization or none.

All six variables are grand-mean centered.

$$\begin{pmatrix} 3.78 & 0.05 & 0.06 & -0.18 & -0.02 & -2.33 & 0.36 \\ 0.05 & 0.23 & -0.04 & -0.10 & 0.03 & 0.50 & 0.04 \\ -0.06 & -0.04 & 0.23 & 0.00 & -0.03 & 0.11 & -0.02 \\ -0.18 & -0.10 & 0.00 & 0.25 & -0.03 & -0.36 & -0.04 \\ -0.02 & 0.03 & -0.02 & -0.03 & 0.07 & -0.19 & 0.01 \\ -1.33 & 0.50 & 0.11 & -0.36 & -0.19 & 35.51 & 0.07 \\ 0.36 & 0.04 & -0.03 & -0.04 & 0.01 & 0.07 & 0.25 \end{pmatrix}$$

References

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. Journal of the American Statistical Association, 113(523):1112–1121.

Nahum-Shani, I. and Almirall, D. (2019). An introduction to adaptive interventions and smart designs in education. ncser 2020-001. National center for special education research.