

Online Appendix to: “The Value of Design Diversity for Knowledge
Accumulation”

Contents

A Proofs	1
B Additional figures	17
C Design artifacts under the classical test theory framework	18
D Joint tests rather than direct aggregation	20
E A Bayesian framework	21
F Imperfectly correlated artifacts under harmonization in the Bayesian framework	24
G Research design as both bias and sampling variance	26
H Within-study design diversity	28
References	34

A Proofs

Proof of Lemma 1. The decision-maker's utility equals $f(\boldsymbol{\tau})$ if she chooses $a = 1$ and zero otherwise. Hence, the maximal utility $u(a^*)$ that the decision-maker could guarantee herself if she knew $f(\boldsymbol{\tau})$ is

$$u(a^*) = \max \{f(\boldsymbol{\tau}), 0\}.$$

Since the decision-maker does not know $f(\boldsymbol{\tau})$, she follows the decision-rule in equation 4 and chooses $a = 1$ if and only if $f(\hat{\boldsymbol{\tau}}) > 0$. The decision-maker's regret is given by

$$r(\boldsymbol{\tau}, \boldsymbol{\delta}(h)) = \max \{f(\boldsymbol{\tau}), 0\} - \Pr [f(\hat{\boldsymbol{\tau}}) > 0] f(\boldsymbol{\tau}). \quad (1)$$

The distribution of treatment effect estimate $\hat{\tau}_i^j$ generated in context i with design j is

$$\hat{\tau}_i^j \sim \mathcal{N}(\tau_i + \delta^j, \sigma^2).$$

Since $f(\cdot)$ is linear, we can write it as

$$f(\mathbf{x}) = \sum_{i=1}^2 b_i x_i,$$

where $b_1 = b_2 = \frac{1}{2}$ yields the cross-context average and $b_1 = 1$ and $b_2 = -1$ the difference. The linearity of $f(\cdot)$ together with the properties of the normal distribution imply that the estimate $f(\hat{\boldsymbol{\tau}})$ is distributed as follows:

$$f(\hat{\boldsymbol{\tau}}) \sim \mathcal{N} \left(f(\boldsymbol{\tau}) + f(\boldsymbol{\delta}(h)), \sigma^2 \sum_{i=1}^2 b_i^2 \right).$$

$f(\hat{\boldsymbol{\tau}})$ exceeds zero with probability

$$\Pr [f(\hat{\boldsymbol{\tau}}) > 0] = \Phi \left(\frac{f(\boldsymbol{\tau}) + f(\boldsymbol{\delta}(h))}{\sigma \sqrt{\sum_{i=1}^2 b_i^2}} \right),$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Plugging this expression into equation 1 and simplifying yields the following expression for the decision-maker's regret:

$$r(\boldsymbol{\tau}, \boldsymbol{\delta}(h)) = \begin{cases} f(\boldsymbol{\tau})\Phi\left(\frac{-f(\boldsymbol{\tau})-f(\boldsymbol{\delta}(h))}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) & \text{if } f(\boldsymbol{\tau}) > 0 \\ -f(\boldsymbol{\tau})\Phi\left(\frac{f(\boldsymbol{\tau})+f(\boldsymbol{\delta}(h))}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) & \text{if } f(\boldsymbol{\tau}) \leq 0. \end{cases} \quad (2)$$

Note that $\sqrt{\sum_{i=1}^2 b_i^2}$ simplifies to $\mathcal{B}_{f(\cdot)=(x_1+x_2)/2} = \frac{1}{\sqrt{2}}$ for the evidence aggregation and to $\mathcal{B}_{f(\cdot)=x_1-x_2} = \sqrt{2}$ for the external validity case. Moreover, the decision-maker's regret decreases in the artifact term $f(\boldsymbol{\delta}(h))$ if $f(\boldsymbol{\tau}) > 0$ but increases in $f(\boldsymbol{\delta}(h))$ if $f(\boldsymbol{\tau}) \leq 0$. Therefore, the worst regret obtains either if $f(\boldsymbol{\tau}) > 0$ and $f(\boldsymbol{\delta}(h))$ takes its smallest possible value or if $f(\boldsymbol{\tau}) \leq 0$ and $f(\boldsymbol{\delta}(h))$ takes its largest possible value. Hence, we can re-write the decision-maker's choice problem as stated in the result.

Note that $\tilde{r}(\boldsymbol{\tau}, h)$ depends on $\boldsymbol{\tau}$ only through $f(\boldsymbol{\tau})$. Hence, we can write $\tilde{r}(\boldsymbol{\tau}, h) := \hat{r}(f(\boldsymbol{\tau}), h)$. Next, we show that $\hat{r}(f(\boldsymbol{\tau}), h)$ is strictly quasi-concave in $f(\boldsymbol{\tau})$ for $f(\boldsymbol{\tau}) > 0$. To do so, we use $\hat{r}_+(f(\boldsymbol{\tau}))$ to denote the function $\hat{r}(f(\boldsymbol{\tau}), h)$ on the positive domain $[0, \infty]$ and find the first derivative $\hat{r}'_+(f(\boldsymbol{\tau}), h)$ of this function using the product rule:

$$\hat{r}'_+(f(\boldsymbol{\tau}), h) = \Phi\left(\frac{-f(\boldsymbol{\tau}) - \boldsymbol{\delta}(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) - \frac{f(\boldsymbol{\tau})}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\phi\left(\frac{-f(\boldsymbol{\tau}) - \boldsymbol{\delta}(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right),$$

where we use $\phi(\cdot)$ to denote the PDF of the standard normal distribution.

Next, consider the behavior of $\hat{r}'_+(f(\boldsymbol{\tau}), h)$ as $f(\boldsymbol{\tau}) \rightarrow 0$. The first term of the sum goes to $\Phi\left(\frac{-\boldsymbol{\delta}(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) > 0$. The second term $\frac{f(\boldsymbol{\tau})}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\phi\left(\frac{-f(\boldsymbol{\tau}) - \boldsymbol{\delta}(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right)$ goes to zero. Hence, $\hat{r}'_+(f(\boldsymbol{\tau}), h) > 0$ and $\hat{r}_+(f(\boldsymbol{\tau}))$ is increasing as $f(\boldsymbol{\tau}) \rightarrow 0$.

Now we turn to the behavior of $\hat{r}'_+(f(\boldsymbol{\tau}), h)$ as $f(\boldsymbol{\tau}) \rightarrow \infty$. To do so, we note that $\Phi(z) \sim \frac{\phi(z)}{-z}$. Plugging this approximation into the first derivative gives

$$\begin{aligned} \hat{r}'_+(f(\boldsymbol{\tau}), h) &= \phi\left(\frac{-f(\boldsymbol{\tau}) - \boldsymbol{\delta}(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) \frac{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}{f(\boldsymbol{\tau}) + \boldsymbol{\delta}(h)} - \frac{f(\boldsymbol{\tau})}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\phi\left(\frac{-f(\boldsymbol{\tau}) - \boldsymbol{\delta}(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) \\ &= \phi\left(\frac{-f(\boldsymbol{\tau}) - \boldsymbol{\delta}(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) \left(\frac{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}{f(\boldsymbol{\tau}) + \boldsymbol{\delta}(h)} - \frac{f(\boldsymbol{\tau})}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right). \end{aligned}$$

Now, it is easy to see that $\hat{r}'_+(f(\boldsymbol{\tau}), h)$ approaches zero from below as $f(\boldsymbol{\tau}) \rightarrow \infty$, since the term in the brackets tends to negative infinity and $\phi\left(\frac{-f(\boldsymbol{\tau})-\delta(h)}{\sigma\sqrt{\sum_{i=1}^2 b_i^2}}\right) \rightarrow 0$. Hence, $\hat{r}'_+(f(\boldsymbol{\tau}), h) < 0$ and $\hat{r}_+(f(\boldsymbol{\tau}))$ is decreasing as $\tau \rightarrow \infty$.

The previous two facts imply that $\hat{r}'_+(f(\boldsymbol{\tau}), h)$ must cross zero at least once on $[0, \infty)$. We will now show that this crossing is unique. To do so, we find the second derivative of $\hat{r}_+(f(\boldsymbol{\tau}))$ w.r.t. $f(\boldsymbol{\tau})$:

$$\hat{r}''_+(f(\boldsymbol{\tau}), h) = \frac{f(\boldsymbol{\tau})(f(\boldsymbol{\tau}) + f(\delta(h))) - 2\sum_{i=1}^2 b_i^2 \sigma}{\sqrt{2\pi}\sigma^3 \left(\sum_{i=1}^2 b_i^2\right)^3} e^{-\frac{(f(\boldsymbol{\tau})+f(\delta(h)))^2}{2\sigma\sum_{i=1}^2 b_i^2}}.$$

This expression is negative whenever $f(\boldsymbol{\tau})(f(\boldsymbol{\tau}) + f(\delta(h))) - 2\sum_{i=1}^2 b_i^2 \sigma < 0$. This expression is a quadratic in $f(\boldsymbol{\tau})$ with two roots:

$$f(\boldsymbol{\tau})_1 = \frac{1}{2}(-f(\delta(h)) - \sqrt{8\sigma^2 \sum_{i=1}^2 b_i^2 + f(\delta(h))^2})$$

$$f(\boldsymbol{\tau})_2 = \frac{1}{2}(-f(\delta(h)) + \sqrt{8\sigma^2 \sum_{i=1}^2 b_i^2 + f(\delta(h))^2})$$

It is easy to show that $f(\boldsymbol{\tau})_1 < 0$ and $f(\boldsymbol{\tau})_2 > 0$. Hence, we have shown that $\hat{r}''_+(f(\boldsymbol{\tau}), h)$ crosses zero only once within $[0, \infty)$. In particular, it is the case that $\hat{r}'_+(f(\boldsymbol{\tau}), h)$ first decreases as long as $0 < f(\boldsymbol{\tau}) \leq f(\boldsymbol{\tau})_2$ and then increases once $f(\boldsymbol{\tau}) > f(\boldsymbol{\tau})_2$. It follows from this and from the fact that $\hat{r}'_+(f(\boldsymbol{\tau}))$ tends to a positive constant as $f(\boldsymbol{\tau}) \rightarrow 0$ and to zero from below as $f(\boldsymbol{\tau}) \rightarrow \infty$ that $\hat{r}'_+(f(\boldsymbol{\tau}), h)$ crosses zero a single time on $[0, \infty)$. This completes the proof of $\hat{r}(f(\boldsymbol{\tau}), h)$ being strictly quasi-concave in $f(\boldsymbol{\tau})$ for $f(\boldsymbol{\tau}) > 0$. By symmetry, it follows that $\hat{r}(f(\boldsymbol{\tau}), h)$ is strictly quasi-concave in $f(\boldsymbol{\tau})$ for $f(\boldsymbol{\tau}) \leq 0$.

We have thus shown that there exists a unique value $\overline{f(\boldsymbol{\tau})} > 0$ such that $\hat{r}(\overline{f(\boldsymbol{\tau})}) = \max_{f(\boldsymbol{\tau})|f(\boldsymbol{\tau})>0} \hat{r}(f(\boldsymbol{\tau}), h)$ and a unique value $\underline{f(\boldsymbol{\tau})} < 0$ such that $\hat{r}(\underline{f(\boldsymbol{\tau})}) = \max_{f(\boldsymbol{\tau})|f(\boldsymbol{\tau})\leq 0} \hat{r}(f(\boldsymbol{\tau}), h)$. Hence, $\max_{f(\boldsymbol{\tau})} \hat{r}(f(\boldsymbol{\tau}), h) = \max\{\hat{r}(\underline{f(\boldsymbol{\tau})}), \hat{r}(\overline{f(\boldsymbol{\tau})})\}$. This maximum is not unique if $\hat{r}(\underline{f(\boldsymbol{\tau})}) = \hat{r}(\overline{f(\boldsymbol{\tau})})$.

Finally, the above steps imply that

$$\arg \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, h) = \begin{cases} \left\{ \boldsymbol{\tau}^* \in \mathbb{R}^2 \mid f(\boldsymbol{\tau}) = \overline{f(\boldsymbol{\tau})} \right\} & \text{if } \hat{r}(f(\boldsymbol{\tau})) < \hat{r}(\overline{f(\boldsymbol{\tau})}) \\ \left\{ \boldsymbol{\tau}^* \in \mathbb{R}^2 \mid f(\boldsymbol{\tau}) \in \left\{ \underline{f(\boldsymbol{\tau})}, \overline{f(\boldsymbol{\tau})} \right\} \right\} & \text{if } \hat{r}(f(\boldsymbol{\tau})) = \hat{r}(\overline{f(\boldsymbol{\tau})}) \\ \left\{ \boldsymbol{\tau}^* \in \mathbb{R}^2 \mid f(\boldsymbol{\tau}) = \underline{f(\boldsymbol{\tau})} \right\} & \text{if } \hat{r}(f(\boldsymbol{\tau})) > \hat{r}(\overline{f(\boldsymbol{\tau})}), \end{cases}$$

where linearity of $f(\cdot)$ implies that $\boldsymbol{\tau}^*$ exists and that it is generally not unique.

Proof of Proposition 1. For the cross-context average as an estimand, we have

$$f(\boldsymbol{\delta}(h)) = \begin{cases} \delta_1 & \text{if } h = 1 \\ \frac{\delta_1 + \delta_2}{2} & \text{if } h = 0. \end{cases}$$

Following equation A, the estimate $f(\hat{\boldsymbol{\tau}}) = \frac{\hat{\tau}_1^j + \hat{\tau}_2^j}{2}$ is thus distributed as follows:

$$\frac{\hat{\tau}_1^j + \hat{\tau}_2^j}{2} \sim \begin{cases} \mathcal{N}\left(\frac{\tau_1 + \tau_2}{2} + \delta_1, \frac{\sigma^2}{2}\right) & \text{if } h = 1 \\ \mathcal{N}\left(\frac{\tau_1 + \tau_2}{2} + \frac{\delta_1 + \delta_2}{2}, \frac{\sigma^2}{2}\right) & \text{if } h = 0. \end{cases}$$

Moreover, we have $f(\boldsymbol{\delta}^{\min}(1)) = \underline{\delta}$ and $f(\boldsymbol{\delta}^{\max}(1)) = \bar{\delta}$. Hence, the decision-maker's maximum regret under harmonization is

$$R(1) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 1) = \begin{cases} \frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{-\frac{\tau_1 + \tau_2}{2} - \delta}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} > 0 \\ -\frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{\frac{\tau_1 + \tau_2}{2} + \bar{\delta}}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} \leq 0. \end{cases} \quad (3)$$

Under research design diversity, we have $f(\boldsymbol{\delta}^{\max}(0)) = \frac{\bar{\delta} + \bar{\delta} - \Delta}{2} = \bar{\delta} - \frac{\Delta}{2}$ and $f(\boldsymbol{\delta}^{\min}(0)) = \frac{\underline{\delta} + \underline{\delta} + \Delta}{2} = \underline{\delta} + \frac{\Delta}{2}$. Hence, the decision-maker's maximum regret is given by

$$R(0) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 0) = \begin{cases} \frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{-\frac{\tau_1 + \tau_2}{2} - \underline{\delta} - \frac{\Delta}{2}}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} > 0 \\ -\frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{\frac{\tau_1 + \tau_2}{2} + \bar{\delta} - \frac{\Delta}{2}}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} \leq 0. \end{cases} \quad (4)$$

Comparing equations 3 and 4, it is immediately obvious that the decision-maker's regret under design

harmonization $\tilde{r}(\boldsymbol{\tau}, 1)$ always exceeds her regret $\tilde{r}(\boldsymbol{\tau}, 0)$ under design diversity.

Proof of Proposition 2. For the cross-context difference as an estimand, we have

$$f(\boldsymbol{\delta}(h)) = \begin{cases} 0 & \text{if } h = 1 \\ \delta_1 - \delta_2 & \text{if } h = 0. \end{cases}$$

and hence the estimate $f(\hat{\boldsymbol{\tau}}) = \hat{\tau}_1^j - \hat{\tau}_2^j$ is distributed as follows:

$$\hat{\tau}_1^j - \hat{\tau}_2^j \sim \begin{cases} \mathcal{N}(\tau_1 - \tau_2, 2\sigma^2) & \text{if } h = 1 \\ \mathcal{N}(\tau_1 - \tau_2 + \delta_1 - \delta_2, 2\sigma^2) & \text{if } h = 0. \end{cases}$$

Because harmonization holds artifacts constant across studies, we have $f(\boldsymbol{\delta}^{\min}(1)) = f(\boldsymbol{\delta}^{\max}(1)) = 0$.

Hence, the decision-maker's regret is

$$R(1) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 1) = \begin{cases} (\tau_1 - \tau_2) \Phi\left(\frac{-(\tau_1 - \tau_2)}{\sqrt{2}\sigma}\right) & \text{if } \tau_1 - \tau_2 > 0 \\ (\tau_1 - \tau_2) \Phi\left(\frac{\tau_1 - \tau_2}{\sqrt{2}\sigma}\right) & \text{if } \tau_1 - \tau_2 \leq 0. \end{cases} \quad (5)$$

Under design diversity, we have $f(\boldsymbol{\delta}^{\min}(0)) = \underline{\delta} - \bar{\delta}$ and $f(\boldsymbol{\delta}^{\max}(0)) = \bar{\delta} - \underline{\delta}$. Hence, the decision-maker's regret is

$$R(0) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 0) = \begin{cases} (\tau_1 - \tau_2) \Phi\left(\frac{-(\tau_1 - \tau_2) - (\underline{\delta} - \bar{\delta})}{\sigma/\sqrt{2}}\right) & \text{if } \tau_1 - \tau_2 > 0 \\ (\tau_1 - \tau_2) \Phi\left(\frac{\tau_1 - \tau_2 + \bar{\delta} - \underline{\delta}}{\sigma/\sqrt{2}}\right) & \text{if } \tau_1 - \tau_2 \leq 0. \end{cases} \quad (6)$$

Harmonization is obviously preferred.

Proof of Proposition 3. Given the new decision-rule, the decision-maker's regret is

$$r(\boldsymbol{\tau}, \boldsymbol{\delta}(h), c) = \max\{f(\boldsymbol{\tau}), 0\} - \Pr[f(\hat{\boldsymbol{\tau}}) > c] f(\boldsymbol{\tau}). \quad (7)$$

Since the estimate distribution remains identical, we have

$$\Pr [f(\hat{\boldsymbol{\tau}}) > c] = \Phi \left(\frac{f(\boldsymbol{\tau}) + f(\boldsymbol{\delta}(h)) - c}{\sigma \sqrt{\sum_{i=1}^2 b_i^2}} \right),$$

and so the decision-maker's regret is

$$r(\boldsymbol{\tau}, \boldsymbol{\delta}(h), c) = \begin{cases} f(\boldsymbol{\tau}) \Phi \left(\frac{c - f(\boldsymbol{\tau}) - f(\boldsymbol{\delta}(h))}{\sigma \sqrt{\sum_{i=1}^2 b_i^2}} \right) & \text{if } f(\boldsymbol{\tau}) > 0 \\ -f(\boldsymbol{\tau}) \Phi \left(\frac{f(\boldsymbol{\tau}) + f(\boldsymbol{\delta}(h)) - c}{\sigma \sqrt{\sum_{i=1}^2 b_i^2}} \right) & \text{if } f(\boldsymbol{\tau}) \leq 0, \end{cases} \quad (8)$$

and, as before, the decision-maker's choice problem can be rewritten as

$$\min_{h,c} R(h, c) \text{ where } R(h, c) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, h, c) = \begin{cases} f(\boldsymbol{\tau}) \Phi \left(\frac{c - f(\boldsymbol{\tau}) - f(\boldsymbol{\delta}^{\min}(h))}{\sigma \sqrt{\sum_{i=1}^2 b_i^2}} \right) & \text{if } f(\boldsymbol{\tau}) > 0 \\ -f(\boldsymbol{\tau}) \Phi \left(\frac{f(\boldsymbol{\tau}) + f(\boldsymbol{\delta}^{\max}(h)) - c}{\sigma \sqrt{\sum_{i=1}^2 b_i^2}} \right) & \text{if } f(\boldsymbol{\tau}) \leq 0. \end{cases} \quad (9)$$

It is easy to show that, for a given c , lemma 1 remains true. It follows from equation 9 that the presence of c does not alter how the decision-maker's maximum regret under harmonization compares to that under research design diversity for a given $f(\cdot)$, as long as the decision-maker uses the same cutoff c in both cases.

Next, we show that the decision-maker's optimal choice of c does not vary with her harmonization choice h . $\tilde{r}(\boldsymbol{\tau}, h, c)$ decreases in c if $f(\boldsymbol{\tau}) \leq 0$ and increases in c if $f(\boldsymbol{\tau}) > 0$. This behavior is intuitive. If $f(\boldsymbol{\tau}) \leq 0$, the right decision is $a^* = 0$. Using a larger cutoff makes it less likely that the decision-maker implements $a = 1$ for a given $f(\boldsymbol{\tau})$ and hence regret decreases in c if $f(\boldsymbol{\tau}) \leq 0$. Conversely, if $f(\boldsymbol{\tau}) > 0$, the optimal decision is $a^* = 1$. Hence, regret increases in c if $f(\boldsymbol{\tau}) > 0$. Thus, increasing c makes the decision-maker worse off if $f(\boldsymbol{\tau}) > 0$ and better off if $f(\boldsymbol{\tau}) \leq 0$. The optimal cutoff obtains when these effects balance each other out.

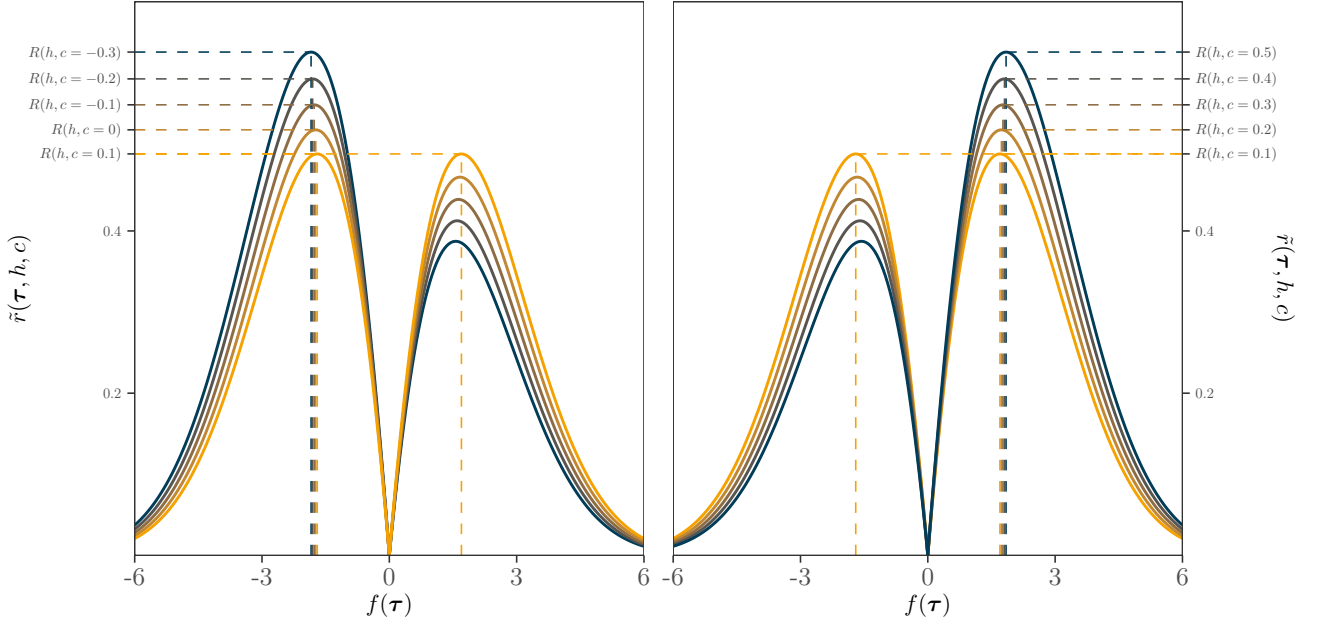


Figure A1: The decision-maker’s regret as a function of $f(\tau)$ for different cutoffs c

Dashed lines indicate the location of $R(h, c) = \max_{\tau} \tilde{r}(\tau, h, c)$. Plotted for $\sigma\sqrt{\sum_{i=1}^2 b_i^2} = 2$, $f(\delta^{\min}(h)) = -0.5$, $f(\delta^{\max}(h)) = 0.7$.

Figure A1 illustrates this intuition. The left panel shows that, in this example, a decision-maker who uses a cutoff of $c = -0.3$ has incentives to increase this cutoff. The reason is that her maximum regret obtains when $f(\tau) \leq 0$ and this maximum regret decreases as she increases c . However, in doing so, the decision-maker also increases the largest regret that she can experience if $f(\tau) > 0$. Once the decision-maker hits $c = 0.1$, the maximum regret for the case where $f(\tau) \leq 0$ equals the maximum regret for the case where $f(\tau) > 0$. As shown in the right panel, the decision-maker has no incentives to increase c beyond this point. Doing so would shift the decision-maker’s maximum possible regret to the case where $f(\tau) > 0$ and increase it beyond that which obtains at $c = 0.1$. Hence, in this example the optimal cutoff is $c^* = 0.1$.

The case in which the worst-case regret for $f(\tau) \leq 0$ equals the worst-case regret for $f(\tau) > 0$ arises whenever the decision-maker choose c^* such that the function $\tilde{r}(\tau, h, c)$ is symmetric around the origin in $f(\tau)$.¹ It is easy to see from equation 9 that this symmetry arises whenever

$$c^* = \frac{f(\delta^{\min}(h)) + f(\delta^{\max}(h))}{2}. \quad (10)$$

¹While difficult to show analytically, this result can easily be verified numerically.

Plugging the values for $f(\boldsymbol{\delta}^{\min}(h))$ and $f(\boldsymbol{\delta}^{\max}(h))$ that we derived for different estimands and harmonization choices in the previous sections into equation 10 yields c^* as given in the proposition.

Proof of Proposition 4. A study in context i with design j now produces an estimate of the effect τ_i given by

$$\hat{\tau}_i^j = \tau_i + \omega\delta^j + (1 - \omega)\eta_i^j + \epsilon_i.$$

Assuming as before and w.l.o.g. that harmonization means using design 1 for both studies, while diversity means using design 1 in context 1 and design 2 in context 2, the choice between diversity and harmonization becomes a choice between the following two vectors $\boldsymbol{\delta}$ of design artifacts:

$$\boldsymbol{\delta}(h) = \begin{cases} (\omega\delta^1 + (1 - \omega)\eta_1^1, \omega\delta^1 + (1 - \omega)\eta_2^1) & \text{if } h = 1, \\ (\omega\delta^1 + (1 - \omega)\eta_1^1, \omega\delta^2 + (1 - \omega)\eta_2^2) & \text{if } h = 0. \end{cases}$$

Let's first consider the evidence aggregation case where $f(\boldsymbol{\tau}) = \frac{\tau_1 + \tau_2}{2}$. Plugging in the worst-case artifacts, it is easy to find the largest possible shift $f(\boldsymbol{\delta}^{\max}(h))$ and the smallest possible shift $f(\boldsymbol{\delta}^{\min}(h))$ of the location of the estimate distribution under harmonization and design diversity:

$$f(\boldsymbol{\delta}^{\max}(h)) = \begin{cases} \omega\bar{\delta} + (1 - \omega)\frac{\bar{\eta}_1 + \bar{\eta}_2}{2} & \text{if } h = 1, \\ \omega(\bar{\delta} - \frac{\Delta}{2}) + (1 - \omega)\frac{\bar{\eta}_1 + \bar{\eta}_2}{2} & \text{if } h = 0, \end{cases}$$

and

$$f(\boldsymbol{\delta}^{\min}(h)) = \begin{cases} \omega\underline{\delta} + (1 - \omega)\frac{\underline{\eta}_1 + \underline{\eta}_2}{2} & \text{if } h = 1, \\ \omega(\underline{\delta} + \frac{\Delta}{2}) + (1 - \omega)\frac{\underline{\eta}_1 + \underline{\eta}_2}{2} & \text{if } h = 0. \end{cases}$$

These expressions show that the largest possible shift of the estimate distribution is bigger and the smallest possible shift is smaller under design harmonization than under design diversity, i.e. $f(\boldsymbol{\delta}^{\max}(1)) \geq f(\boldsymbol{\delta}^{\max}(0))$ and $f(\boldsymbol{\delta}^{\min}(1)) \leq f(\boldsymbol{\delta}^{\min}(0))$. These inequalities are strict as long as $\omega > 0$. The result follows directly from these facts in combination with equation (7) in the main text.

Next, consider the external validity case where $f(\boldsymbol{\tau}) = \tau_1 - \tau_2$. The largest and smallest possible

shifts of the location of the estimate distribution are given by

$$f(\boldsymbol{\delta}^{max}(h)) = \begin{cases} (1 - \omega) (\bar{\eta}_1 - \underline{\eta}_2) & \text{if } h = 1, \\ \omega (\bar{\delta} - \underline{\delta}) + (1 - \omega) (\bar{\eta}_1 - \underline{\eta}_2) & \text{if } h = 0, \end{cases}$$

and

$$f(\boldsymbol{\delta}^{min}(h)) = \begin{cases} (1 - \omega) (\underline{\eta}_1 - \bar{\eta}_2) & \text{if } h = 1, \\ \omega (\underline{\delta} - \bar{\delta}) + (1 - \omega) (\underline{\eta}_1 - \bar{\eta}_2) & \text{if } h = 0. \end{cases}$$

Here, the largest possible shift of the estimate distribution is bigger and the smallest possible shift is smaller under design diversity than under harmonization, i.e. $f(\boldsymbol{\delta}^{max}(0)) \geq f(\boldsymbol{\delta}^{max}(1))$ and $f(\boldsymbol{\delta}^{min}(0)) \leq f(\boldsymbol{\delta}^{min}(1))$. As before, these inequalities are strict as long as $\omega > 0$. Again, the result follows directly from these facts in combination with equation (7) in the main text.

Proof of Proposition 5. With $N > 2$ contexts, the decision-maker's regret still takes the form of equation (2) in the main text and her decision-problem can still be written as in equation (7), but the vectors $\boldsymbol{\tau}$, $\boldsymbol{\delta}(h)$, $\boldsymbol{\delta}^{max}(h)$, and $\boldsymbol{\delta}^{min}(h)$ are now all N -dimensional, storing one treatment effect or (worst-case) design artifact for each of the N contexts, and the sum in the denominators of the two expressions is taken over all N contexts. The variance of the estimates $f(\hat{\boldsymbol{\tau}})$ is now given by $\sigma^2 \frac{1}{N}$ in the evidence aggregation and by $\sigma^2 \frac{N}{N-1}$ in the external validity case, but, as before, does not impact the decision-maker's choice between diversity and harmonization.

Let's first consider the evidence aggregation case where $f(\boldsymbol{\tau}) = \frac{\sum_{i=1}^N \tau_i}{N}$. Harmonization introduces design artifact δ^1 in all contexts. Hence, the worst-case regret obtains if $\frac{\sum_{i=1}^N \tau_i}{N} \leq 0$ and design 1 induces the largest possible artifact $\bar{\delta}$ in all N studies or if $\frac{\sum_{i=1}^N \tau_i}{N} > 0$ and design 1 induces the smallest possible artifact $\underline{\delta}$ in all N studies, i.e., $f(\boldsymbol{\delta}^{min}(1)) = \underline{\delta}$ and $f(\boldsymbol{\delta}^{max}(1)) = \bar{\delta}$.

If the decision-maker diversifies, she uses design 1 in $n^{(1)}$ contexts and design 2 in the remaining $N - n^{(1)}$ contexts. Suppose first that $n^{(1)} \leq \frac{N}{2}$. Since the decision-maker uses design 2 in weakly more contexts than design 1, the largest possible shift in the estimate distribution occurs if $\delta^2 = \bar{\delta}$ and $\delta^1 = \bar{\delta} - \Delta$, i.e., $f(\boldsymbol{\delta}^{max}(0)) = \frac{(N - n^{(1)})\bar{\delta} + n^{(1)}(\bar{\delta} - \Delta)}{N} = \bar{\delta} - \frac{n^{(1)}}{N}\Delta$. By the same logic, the smallest possible shift in the estimate distribution occurs if $\delta^2 = \underline{\delta}$ and $\delta^1 = \underline{\delta} + \Delta$, i.e., $f(\boldsymbol{\delta}^{min}(0)) =$

$\frac{(N-n^{(1)})\underline{\delta}+n^{(1)}(\underline{\delta}+\Delta)}{N} = \underline{\delta} + \frac{n^{(1)}}{N}\Delta$. The case in which $n^{(1)} > \frac{N}{2}$ is analogous. In summary:

$$f(\boldsymbol{\delta}^{\min}(h)) = \begin{cases} \underline{\delta} & \text{if } h = 1 \\ \underline{\delta} + \frac{n^{(1)}}{N}\Delta & \text{if } h = 0 \text{ and } n^{(1)} \leq \frac{N}{2} \\ \underline{\delta} + \frac{N-n^{(1)}}{N}\Delta & \text{if } h = 0 \text{ and } n^{(1)} > \frac{N}{2}, \end{cases} \quad (11)$$

and

$$f(\boldsymbol{\delta}^{\max}(h)) = \begin{cases} \bar{\delta} & \text{if } h = 1 \\ \bar{\delta} - \frac{n^{(1)}}{N}\Delta & \text{if } h = 0 \text{ and } n^{(1)} \leq \frac{N}{2} \\ \bar{\delta} - \frac{N-n^{(1)}}{N}\Delta & \text{if } h = 0 \text{ and } n^{(1)} > \frac{N}{2}. \end{cases} \quad (12)$$

Regardless of the number $n^{(1)}$ of contexts in which the decision-maker uses design 1 under diversification, the largest possible shift of the estimate distribution remains larger and the smallest possible shift remains smaller under design harmonization than under diversity: $f(\boldsymbol{\delta}^{\max}(1)) > f(\boldsymbol{\delta}^{\max}(0))$ and $f(\boldsymbol{\delta}^{\min}(1)) < f(\boldsymbol{\delta}^{\min}(0))$. Hence, the decision-maker's regret under harmonization always exceeds her regret under diversity and so the decision-maker finds it optimal to diversify.

As long as $n^{(1)} \leq \frac{N}{2}$, $f(\boldsymbol{\delta}^{\max}(0))$ decreases and $f(\boldsymbol{\delta}^{\min}(0))$ increases in $n^{(1)}$. Once $n^{(1)} > \frac{N}{2}$, $f(\boldsymbol{\delta}^{\max}(0))$ increases and $f(\boldsymbol{\delta}^{\min}(0))$ decreases with $n^{(1)}$. Hence, the decision-maker can minimize her maximal regret under design diversity by setting $n^{(1)*} = \frac{N}{2}$.

Turning to the first of the two external validity estimands, suppose for concreteness and w.l.o.g. that the decision-maker's estimand is the difference between the treatment effect in context 1 and all other contexts: $f(\boldsymbol{\tau}) = \tau_1 - \frac{\sum_{i=2}^N \tau_i}{N-1}$. Moreover, assume the decision-maker uses design 1 in context 1 as well as in $n^{(1)} - 1$ other contexts, and design 2 in the remaining $N - n^{(1)}$ contexts if she diversifies.

In this case, the largest and smallest possible shifts of the estimate distribution are as follows:

$$f(\boldsymbol{\delta}^{\min}(h)) = \begin{cases} 0 & \text{if } h = 1 \\ \underline{\delta} - \frac{(n^{(1)}-1)\underline{\delta}+(N-n^{(1)})\bar{\delta}}{N-1} & \text{if } h = 0 \end{cases} \quad (13)$$

and

$$f(\boldsymbol{\delta}^{\max}(h)) = \begin{cases} 0 & \text{if } h = 1 \\ \bar{\delta} - \frac{(n^{(1)}-1)\bar{\delta} + (N-n^{(1)})\underline{\delta}}{N-1} & \text{if } h = 0. \end{cases} \quad (14)$$

As before, if the decision-maker harmonizes, design artifacts difference out and the estimate distribution is centered on the quantity of interest, i.e., $f(\boldsymbol{\delta}^{\min}(1)) = f(\boldsymbol{\delta}^{\max}(1)) = 0$. The same will not be true if the decision-maker diversifies, since $f(\boldsymbol{\delta}^{\min}(0)) < 0$ and $f(\boldsymbol{\delta}^{\max}(0)) > 0$. Hence, the decision-maker prefers harmonization. The case where the decision-maker uses design 2 in context 1 and in $N - n^{(1)} - 1$ other contexts and design 1 in the remaining $n^{(1)}$ contexts is analogous.

Finally, consider the second external validity estimand: $f(\boldsymbol{\tau}) = \frac{\sum_{i=1}^{n_X} \tau_i}{n_X} - \frac{\sum_{i=n_X+1}^N \tau_i}{N-n_X}$. It is straightforward that harmonization, as for the previous estimand, leads design artifacts to cancel out such that the estimate distribution is always centered on the estimand of interest. If the decision-maker diversifies, the shift of the estimate distribution is given by:

$$\begin{aligned} f(\boldsymbol{\delta}(0)) &= [\xi_X \delta^1 + (1 - \xi_X) \delta^2] - [\xi_Y \delta^1 + (1 - \xi_Y) \delta^2] \\ &= (\delta^1 - \delta^2) (\xi_X - \xi_Y). \end{aligned}$$

It follows that the largest and smallest possible shifts of the estimate distribution given the decision-maker's harmonization choice are given by:

$$f(\boldsymbol{\delta}^{\min}(h)) = \begin{cases} 0 & \text{if } h = 1 \\ 0 & \text{if } h = 0 \text{ and } \xi_X = \xi_Y \\ \underline{\delta} - \bar{\delta} & \text{if } h = 0 \text{ and } \xi_X > \xi_Y \\ \bar{\delta} - \underline{\delta} & \text{if } h = 0 \text{ and } \xi_X < \xi_Y \end{cases} \quad (15)$$

and

$$f(\boldsymbol{\delta}^{\max}(h)) = \begin{cases} 0 & \text{if } h = 1 \\ 0 & \text{if } h = 0 \text{ and } \xi_X = \xi_Y \\ \bar{\delta} - \underline{\delta} & \text{if } h = 0 \text{ and } \xi_X > \xi_Y \\ \underline{\delta} - \bar{\delta} & \text{if } h = 0 \text{ and } \xi_X < \xi_Y. \end{cases} \quad (16)$$

It is immediately obvious that $\xi_X = \xi_Y$ implies $f(\boldsymbol{\delta}(0)) = f(\boldsymbol{\delta}^{\max}(0)) = f(\boldsymbol{\delta}^{\min}(0)) = 0$, which implies that $R(0) = R(1)$, i.e., the decision-maker is indifferent between harmonization and diversity. Since the decision-maker uses design 1 in $n^{(1)}$ contexts, this case requires $\xi_X = \xi_Y = \frac{n^{(1)}}{N}$ and that n_X and n_Y are divisible by $\frac{N}{n^{(1)}}$. Moreover, it is easy to see from the expressions below that $f(\boldsymbol{\delta}^{\min}(0)) < 0$ and $f(\boldsymbol{\delta}^{\max}(0)) > 0$ when $\xi_X \neq \xi_Y$. Hence, in this case $R(0) < R(1)$, and the decision-maker prefers harmonization.

Proof of Proposition 6. The decision to harmonize now entails not a choice of a vector of design artifacts but of a vector $\boldsymbol{\sigma}^2(h)$ of variances, where the first element refers to the variance of ϵ_1 and the second element to the variance of ϵ_2 :

$$\boldsymbol{\sigma}^2(h) = \begin{cases} (\sigma^{(1)2}, \sigma^{(1)2}) & \text{if } h = 1 \\ (\sigma^{(1)2}, \sigma^{(2)2}) & \text{if } h = 0 \end{cases}$$

The decision-maker faces the following decision-problem:

$$\min_h R(h), \text{ where } R(h) := \max_{\boldsymbol{\tau}, \boldsymbol{\sigma}^2(h)} r(\boldsymbol{\tau}, \boldsymbol{\sigma}^2(h)). \quad (17)$$

Equation 2 implies the decision-maker's regret increases in the variance of the estimate distribution. Hence, the decision-maker seeks to minimize the largest possible variance. Under harmonization, this worst-case obtains if design 1 has the largest possible variance $\bar{\sigma}^2$. Hence, the decision-maker's regret,

for any estimand $f(\boldsymbol{\tau})$, is

$$R(1) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 1) = \begin{cases} f(\boldsymbol{\tau})\Phi\left(\frac{-f(\boldsymbol{\tau})}{\bar{\sigma}\sqrt{\sum_{i=1}^2 b_i^2}}\right) & \text{if } f(\boldsymbol{\tau}) > 0 \\ -f(\boldsymbol{\tau})\Phi\left(\frac{f(\boldsymbol{\tau})}{\bar{\sigma}\sqrt{\sum_{i=1}^2 b_i^2}}\right) & \text{if } f(\boldsymbol{\tau}) \leq 0. \end{cases} \quad (18)$$

The worst case under design diversity arises if one research design induces the largest possible variance $\bar{\sigma}^2$ and the other one a variance of $\bar{\sigma}^2 - \Delta_\sigma$. In case where the estimand $f(\boldsymbol{\tau})$ is such that one study receives a different weight b_i than the other, the worst cases arises if the study with the greater weight – in absolute value – receives the design with the greater variance. The decision-maker’s regret in this case is given by

$$R(0) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 0) = \begin{cases} f(\boldsymbol{\tau})\Phi\left(\frac{-f(\boldsymbol{\tau})}{\sqrt{\bar{\sigma}^2 \sum_{i=1}^2 b_i^2 - \min\{b_1, b_2\}^2 \Delta_\sigma}}\right) & \text{if } f(\boldsymbol{\tau}) > 0 \\ -f(\boldsymbol{\tau})\Phi\left(\frac{f(\boldsymbol{\tau})}{\sqrt{\bar{\sigma}^2 \sum_{i=1}^2 b_i^2 - \min\{b_1, b_2\}^2 \Delta_\sigma}}\right) & \text{if } f(\boldsymbol{\tau}) \leq 0. \end{cases} \quad (19)$$

It follows that the decision-maker always prefers design diversity.

Proof of Lemma A1. Consider first a decision-maker with the utility function given in equation 29. The decision-maker will choose $a = 1$ whenever her posterior mean belief $m(h) = \mathbb{E}[\theta \mid \hat{\boldsymbol{\tau}}]$ about θ after seeing the results of the two studies exceeds her prior $\mathbb{E}[\theta]$ about this estimand. Hence, prior to seeing the study estimates, the decision-maker’s expected utility is given by

$$\begin{aligned} V(h) &= \Pr(m(h) > \mathbb{E}[\theta]) \mathbb{E}[\theta \mid m(h) > \mathbb{E}[\theta]] \\ &\quad + \Pr(m(h) \leq \mathbb{E}[\theta]) \mathbb{E}[\theta] \end{aligned} \quad (20)$$

Since the decision-maker decides whether to harmonize *prior* to seeing the study estimates, $m(h)$ is a random variable at the time of decision-making. Standard results (Schlaifer and Raiffa, 1961) imply that

$$m(h) \sim \mathcal{N}(\mathbb{E}[\theta], v_{\text{prior}} - v_{\text{post}}(h)), \quad (21)$$

where v_{prior} is the variance of the decision-maker’s prior belief about θ .

It follows that $m(h)$ exceeds $\mathbb{E}[\theta]$ with probability $\frac{1}{2}$. Moreover,

$$\mathbb{E}[\theta \mid m(h) > \mathbb{E}[\theta]] = \mathbb{E}_{\hat{\tau}}[m(h) \mid m(h) > \mathbb{E}[\theta]],$$

where $\mathbb{E}_{\hat{\tau}}[\cdot]$ is taken only over the possible estimates $\hat{\tau}$. Substituting these quantities into equation 20 and relying on the properties of a truncated normal distribution, we can now write the decision-maker's expected utility as follows:

$$\begin{aligned} V(h) &= \frac{1}{2} \left(\mathbb{E}[\theta] + \sqrt{v_{prior} - v_{post}(h)} \sqrt{\frac{2}{\pi}} \right) + \frac{1}{2} \mathbb{E}[\theta] \\ &= \mathbb{E}[\theta] \sqrt{\frac{v_{prior} - v_{post}(h)}{2\pi}}. \end{aligned} \tag{22}$$

Equation (22) implies that $V(1) > V(0)$ if and only if $v_{post}(1) < v_{post}(0)$.

Second, consider a decision-maker with the utility function given in equation 30. DeGroot (2005, chap. 11.2) proves that the decision-maker's expected loss under a squared error loss function is equal to the expected posterior variance. To see why, note that the decision-maker's optimal choice e^* after observing $\hat{\tau}$ is $e^* = \mathbb{E}[\theta \mid \hat{\tau}] = m(h)$. Hence the decision-maker's expected utility is

$$V(h) = -\mathbb{E} \left[(\theta - \mathbb{E}[\theta \mid \hat{\tau}])^2 \right] = -\mathbb{E}[v_{post}(h)] = -v_{post}(h).$$

The last equality follows from the fact that the posterior variance does not depend on the realization of $\hat{\tau}$. Hence, again, $V(1) > V(0)$ if and only if $v_{post}(1) < v_{post}(0)$.

Proof of Proposition A1. First, let $\theta = \tau$. We derive the variance, $v_{post}(h)$, of the decision-maker's belief about τ using theorem B.7 in Greene (2011). Specifically, let $\mathbf{x}_1 = \tau$ and $\mathbf{x}_2 = \hat{\tau}$. The variance-covariance matrix of \mathbf{x}_1 and \mathbf{x}_2 under harmonization is given by

$$\Sigma(1) = \begin{bmatrix} v_{\tau} & v_{\tau} & v_{\tau} \\ v_{\tau} & v_{\tau} + v + v_{\delta} + \sigma^2 & v_{\tau} + v_{\delta} \\ v_{\tau} & v_{\tau} + v_{\delta} & v_{\tau} + v + v_{\delta} + \sigma^2 \end{bmatrix}.$$

The variance-covariance matrix of \mathbf{x}_1 and \mathbf{x}_2 under design diversity is given by

$$\boldsymbol{\Sigma}(0) = \begin{bmatrix} v_\tau & v_\tau & v_\tau \\ v_\tau & v_\tau + v + v_\delta + \sigma^2 & v_\tau \\ v_\tau & v_\tau & v_\tau + v + v_\delta + \sigma^2 \end{bmatrix}.$$

The theorem implies that the posterior variance under harmonization is given by

$$v_{post}(h) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}(h))^{-1}\boldsymbol{\Sigma}_{21}, \quad (23)$$

where it follows from $\boldsymbol{\Sigma}(h)$ that

$$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} v_\tau & v_\tau \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{11} = v_\tau,$$

$$\boldsymbol{\Sigma}_{21} = \begin{bmatrix} v_\tau \\ v_\tau \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{22}(1) = \begin{bmatrix} v_\tau + v + v_\delta + \sigma^2 & v_\tau + v_\delta \\ v_\tau + v_\delta & v_\tau + v + v_\delta + \sigma^2 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma}_{22}(0) = \begin{bmatrix} v_\tau + v + v_\delta + \sigma^2 & v_\tau \\ v_\tau & v_\tau + v + v_\delta + \sigma^2 \end{bmatrix}.$$

Simplifying equation 23 yields

$$v_{post}(h) = \begin{cases} \frac{(\sigma^2 + v + 2v_\delta)v_\tau}{\sigma^2 + v + 2(v_\delta + v_\tau)} & \text{if } h = 1 \\ \frac{(\sigma^2 + v + v_\delta)v_\tau}{\sigma^2 + v + v_\delta + 2v_\tau} & \text{if } h = 0 \end{cases} \quad (24)$$

Equation 24 implies that $v_{post}(1) > v_{post}(0)$.

Second, let $\theta = \frac{\tau_1 + \tau_2}{2}$. We again derive the variance, $v_{post}(h)$, of the decision-maker's belief about $\frac{\tau_1 + \tau_2}{2}$ using theorem B.7 in Greene (2011). Specifically, let $\mathbf{x}_1 = \frac{\tau_1 + \tau_2}{2}$ and $\mathbf{x}_2 = \hat{\tau}$. The

variance-covariance matrix of \mathbf{x}_1 and \mathbf{x}_2 under harmonization is given by

$$\Sigma(1) = \begin{bmatrix} v_\tau + \frac{v}{2} & v_\tau + \frac{v}{2} & v_\tau + \frac{v}{2} \\ v_\tau + \frac{v}{2} & v_\tau + v + v_\delta + \sigma^2 & v_\tau + v_\delta \\ v_\tau + \frac{v}{2} & v_\tau + v_\delta & v_\tau + v + v_\delta + \sigma^2 \end{bmatrix}.$$

The variance-covariance matrix of \mathbf{x}_1 and \mathbf{x}_2 under design diversity is given by

$$\Sigma(0) = \begin{bmatrix} v_\tau + \frac{v}{2} & v_\tau + \frac{v}{2} & v_\tau + \frac{v}{2} \\ v_\tau + \frac{v}{2} & v_\tau + v + v_\delta + \sigma^2 & v_\tau \\ v_\tau + \frac{v}{2} & v_\tau & v_\tau + v + v_\delta + \sigma^2 \end{bmatrix}.$$

Plugging the relevant sub-matrices into equation 23 and simplifying yields

$$v_{post}(h) = \begin{cases} \frac{(\sigma^2 + 2v_\delta)(v + 2v_\tau)}{2(\sigma^2 + v + 2(v_\delta + v_\tau))} & \text{if } h = 1 \\ \frac{(\sigma^2 + v_\delta)(v + 2v_\tau)}{2(\sigma^2 + v + v_\delta + 2v_\tau)} & \text{if } h = 0 \end{cases} \quad (25)$$

Equation 25 implies that $v_{post}(1) > v_{post}(0)$.

Proof of Proposition A2. Let $\theta = \tau_1 - \tau_2$. We again derive the variance, $v_{post}(h)$, of the decision-maker's belief about $\frac{\tau_1 + \tau_2}{2}$ using theorem B.7 in Greene (2011). Specifically, let $\mathbf{x}_1 = \tau_1 - \tau_2$ and $\mathbf{x}_2 = \hat{\tau}_1 - \hat{\tau}_2$. The variance-covariance matrix of \mathbf{x}_1 and \mathbf{x}_2 under harmonization is given by

$$\Sigma(1) = \begin{bmatrix} 2v & 2v \\ 2v & 2(v + \sigma^2) \end{bmatrix}.$$

The variance-covariance matrix of \mathbf{x}_1 and \mathbf{x}_2 under design diversity is given by

$$\Sigma(0) = \begin{bmatrix} 2v & 2v \\ 2v & 2(v + \sigma^2 + v_\delta) \end{bmatrix}.$$

Plugging the relevant sub-matrices into equation 23 and simplifying yields

$$v_{post}(h) = \begin{cases} \frac{2\sigma^2 v}{\sigma^2 + v} & \text{if } h = 1 \\ \frac{2v(\sigma^2 + v_\delta)}{\sigma^2 + v + v_\delta} & \text{if } h = 0 \end{cases} \quad (26)$$

Equation 26 implies that $v_{post}(0) > v_{post}(1)$.

B Additional figures

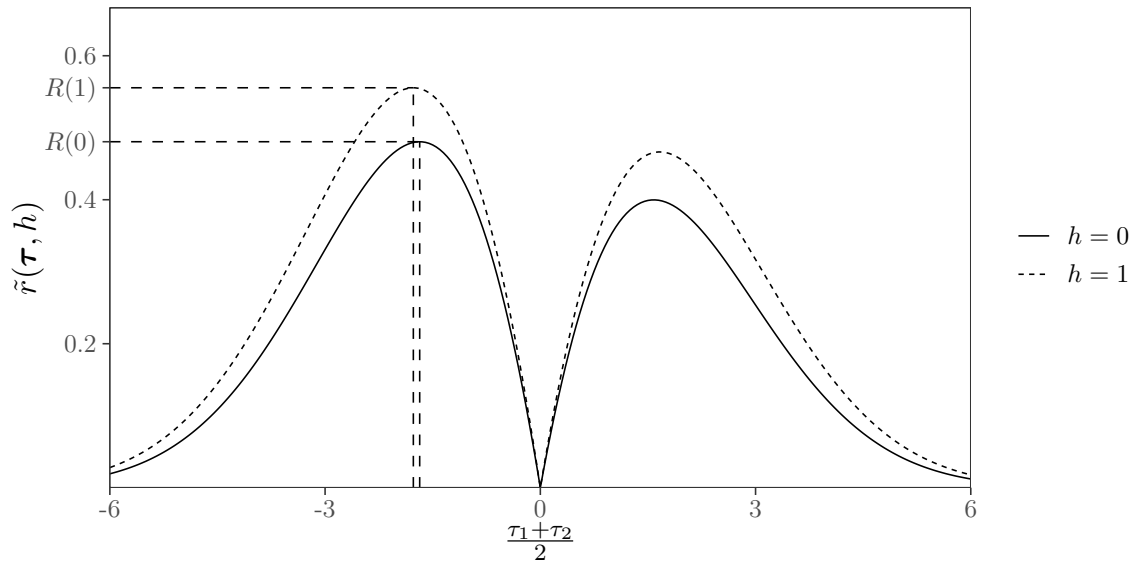


Figure A2: The decision-maker's regret as a function of the cross-context average effect

Dashed lines indicate the location of $R(h) = \max_{\tau} \tilde{r}(\tau, h)$. Note that $\tilde{r}(\tau, h)$ depends on τ only through $f(\tau) = \frac{\tau_1 + \tau_2}{2}$. Plotted for $\sigma^2 = 2$, $\underline{\delta} = -0.5$, $\bar{\delta} = 0.8$.

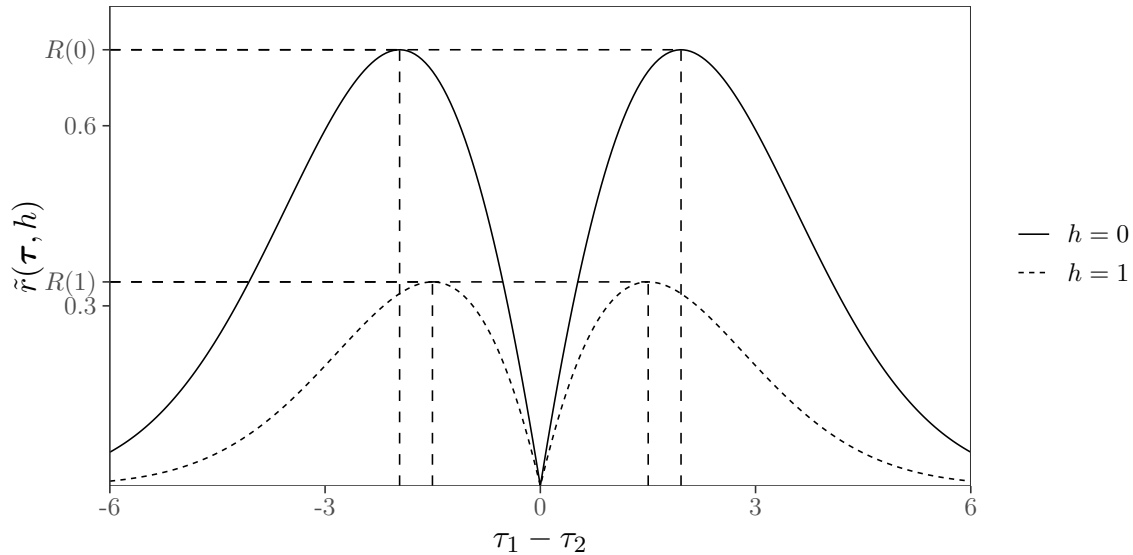


Figure A3: The decision-maker’s regret as a function of the cross-context difference in effects. Dashed lines indicate the location of $R(h) = \max_{\tau} \tilde{r}(\tau, h)$. Note that $\tilde{r}(\tau, h)$ depends on τ only through $f(\tau) = \tau_1 - \tau_2$. Plotted for $\sigma^2 = 8$, $\underline{\delta} = -0.5$, $\bar{\delta} = 0.8$, $\Delta = 0.5$.

C Design artifacts under the classical test theory framework

How can we think about design artifacts being generated on the unit-level? The idea that a test or an item captures an underlying concept with error is key to the literature on measurement. As an example, we here provide one micro-foundation of design artifacts using a measurement error framework in the spirit of classical test theory (Lord and Novick, 2008).

Let us focus on a single context i and index the n_i units in this context by l with $l = 1, \dots, n_i$. Let $z_{il} \in \{0, 1\}$ denote the treatment status of unit l in context i . Invoking the stable unit treatment value assumption (SUTVA), we write unit l ’s potential outcomes as $y_{il}(z_{il})$. The average treatment effect (ATE) in context i is

$$\tau_i = \mathbb{E}[y_{il}(1) - y_{il}(0)] = \sum_{l=1}^{n_i} y_{il}(1) - y_{il}(0).$$

Assume that potential outcomes can never be directly observed but can be *measured*. When using outcome measure j , the *measurable* potential outcomes are given by

$$\tilde{y}_{il}^j = \lambda^j y_{il}(z_{il}) + \nu_{il}^j(z_{il}),$$

where $\lambda^j \in \mathbb{R}$ is a measure-specific scaling factor and $\nu_{il}^j(z_{il}) \in \mathbb{R}$ a measurement error that may vary across units and with a unit's treatment status. A classical way to conceptualize measurement validity in this framework is the correlation $\rho\left(y_{il}(z_{il}), \tilde{y}_{il}^j(z_{il})\right)$ between actual and measurable potential outcomes. Given outcome measure j , the *measurable* ATE is

$$\tilde{\tau}_i^j = \mathbb{E}\left[\tilde{y}_{il}^j(1) - \tilde{y}_{il}^j(0)\right] = \lambda^j \tau_i + \mathbb{E}\left[\nu_{il}^j(1) - \nu_{il}^j(0)\right].$$

Let's say the decision-maker uses a difference-in-means estimator to estimate τ_i . We can then write the difference-in-means estimate that the decision-maker obtains as

$$\hat{\tau}_i^j = \tau_i + \underbrace{\left[(\lambda^j \tau_i - \tau_i) + \mathbb{E}\left[\nu_{il}^j(1) - \nu_{il}^j(0)\right]\right]}_{\delta_i^j} + \epsilon_i,$$

where ϵ_i is an error term that is approximately normally distributed given the finite population CLT.

For purposes of exposition, our main analysis assumes that artifacts of the same design are constant across contexts, i.e., $\delta_1^j = \delta_2^j = \delta^j$. The fairly simple measurement error model presented in this section shows this assumption is restrictive. First, if $\lambda^j \neq 0$, artifacts may differ across contexts because contexts may differ in terms of treatment effects. Second, the errors $\nu_{il}^j(z_{il})$ may vary across contexts for different reasons. First, $\nu_{il}^j(z_{il})$ may vary with context-level features. For example, cultural norms may create social desirability bias in some contexts but not others. Second, $\nu_{il}^j(z_{il})$ may vary with individual-level characteristics – say, social desirability bias is more severe among educated individuals – and the distribution of these characteristics may vary across contexts.

We stress that we do not make this assumption because it is realistic but to make the contrast between design diversity and harmonization stark. In section 5.4 of the main text, we show that our results generalize to a setting in which harmonization introduces an imperfect correlation across design artifacts rather than holding them constant. What is required for our results to be relevant to real world contexts is not that artifacts of the same design are identical across contexts but that using the same operationalizations across studies induces *some* cross-context dependence in artifacts. In other words, using a given design j across contexts needs to produce artifacts δ_1^j and δ_2^j that are closer together than the artifacts $\delta_1^{j'}$ and $\delta_2^{j'}$ that would result if we used design j in context 1 and design j' in context 2.

D Joint tests rather than direct aggregation

Assume $f(\boldsymbol{\tau}) = \mathbb{1}\{\tau_1 > 0 \wedge \tau_2 > 0\}A - \mathbb{1}\{\tau_1 < 0 \vee \tau_2 < 0\}B$, where $A > 0$ and $B > 0$. Given the utility function in equation 2, this specification of $f(\boldsymbol{\tau})$ implies the decision-maker wants to choose $a = 1$ if and only if $\tau_1 > 0$ and $\tau_2 > 0$. To aid her choice a , the decision-maker conducts an intersection-union test of the composite null hypothesis

$$H_0 : H_{0,1} \cup H_{0,2}, \text{ for which } H_{0,1} : \tau_1 \leq 0, H_{0,2} : \tau_2 \leq 0.$$

against the alternative

$$H_A : H_{1,1} \cap H_{1,2}, \text{ for which } H_{1,1} : \tau_1 > 0, H_{1,2} : \tau_2 > 0.$$

The decision-maker rejects the composite null hypothesis only if both individual tests reject the respective component null hypothesis, i.e., yield a p -value that is smaller than the desired type-I error rate α . To conduct this test, the decision-maker constructs the following test statistics

$$Z_1^j = \frac{\hat{\tau}_1^j}{\hat{\sigma}} \text{ and } Z_2^j = \frac{\hat{\tau}_2^j}{\hat{\sigma}},$$

where $\hat{\sigma}$ is the decision-maker's consistent estimate of the standard deviation σ of the study estimates. The decision-maker rejects the composite null hypothesis and implements $a = 1$ if and only if $Z_1^j > z_\alpha$ and $Z_2^j > z_\alpha$, where z_α is the α critical value for the standard normal distribution. In the absence of design artifacts, this procedure yields an asymptotically valid test. In particular, [Berger and Hsu \(1996, Thm. 1\)](#) show that this test has size (i.e., Type I error rate) α if the component tests have size α . Of course, the decision-maker in our case has to contend with the possibility of design artifacts.

Since the two tests are independent, the probability that the decision-maker chooses $a = 1$ equals $\Pr [Z_1^j > z_\alpha] \times \Pr [Z_2^j > z_\alpha]$. We consider the sample-asymptotic properties of the joint test. The asymptotic probability that $Z_i^j > z_\alpha$ is $\Phi(\frac{\tau_i + \delta_i^j}{\sigma} - z_\alpha)$. It follows that the decision-maker's regret

under harmonization is given by

$$r(\boldsymbol{\tau}, \boldsymbol{\delta}(1)) = \begin{cases} A \left(1 - \Phi \left(\frac{\tau_1 + \delta^1}{\sigma} - z_\alpha \right) \Phi \left(\frac{\tau_2 + \delta^1}{\sigma} - z_\alpha \right) \right) & \text{if } \tau_1 > 0 \wedge \tau_2 > 0 \\ B \Phi \left(\frac{\tau_1 + \delta^1}{\sigma} - z_\alpha \right) \Phi \left(\frac{\tau_2 + \delta^1}{\sigma} - z_\alpha \right) & \text{if } \tau_1 \leq 0 \vee \tau_2 \leq 0, \end{cases} \quad (27)$$

while regret under diversity is given by

$$r(\boldsymbol{\tau}, \boldsymbol{\delta}(0)) = \begin{cases} A \left(1 - \Phi \left(\frac{\tau_1 + \delta^1}{\sigma} - z_\alpha \right) \Phi \left(\frac{\tau_2 + \delta^2}{\sigma} - z_\alpha \right) \right) & \text{if } \tau_1 > 0 \wedge \tau_2 > 0 \\ B \Phi \left(\frac{\tau_1 + \delta^1}{\sigma} - z_\alpha \right) \Phi \left(\frac{\tau_2 + \delta^2}{\sigma} - z_\alpha \right) & \text{if } \tau_1 \leq 0 \vee \tau_2 \leq 0. \end{cases} \quad (28)$$

It is straightforward to see from equations 27 and 28 that the decision-maker will prefer research design diversity over harmonization. The reasoning is identical to the case of evidence aggregation in the form of the cross-context average effect. Research design diversity guards against the worst case of two estimates that are heavily biased in the same direction.

In the two-context case that we consider here, the intersection union test is equivalent to the partial conjunction test in Egami and Hartman (2023). For situations with three or more contexts, the difference between the two tests becomes meaningful. The partial conjunction test pertains to a hypothesis about the share of contexts in which the effect has a given sign. Hence, in this case, the decision-maker’s regret would be defined in terms of all the different combinations of estimates that could lead one to reject the null hypothesis. The resulting expressions would resemble those in equations 27 and 28 but would take into account the set of possible estimate combinations for which the decision-maker would reject the null hypothesis. In other words, our conclusion that research design diversity is beneficial extends to aggregation through the partial conjunction test as well.

E A Bayesian framework

Consider a superpopulation of contexts where context-level treatment effects τ_i are drawn from a common normal distribution

$$\tau_i \sim \mathcal{N}(\tau, v),$$

with mean $\tau \in \mathbb{R}$ and variance $v > 0$. In meta-analyses, researchers are typically interested in learning about τ , but may also learn about v . Since a framework with unknown v does not allow for closed-form solutions, we assume, for simplicity, that v is known and focus on inferences about treatment

effects as in our minimax regret framework. The decision-maker observes neither the average effect τ nor context-specific effects τ_i . Her prior beliefs about τ are given by $\mathcal{N}(\mu, v_\tau)$, where $\mu \in \mathbb{R}$ is the mean and $v_\tau > 0$ the variance.

Assume w.l.o.g. that the decision-maker runs her two studies in contexts 1 and 2. To align the analysis with our main results, we focus on the same estimands – the average $\frac{\tau_1 + \tau_2}{2}$ and the difference $\tau_1 - \tau_2$ in treatment effects across studied contexts. We also consider the population average effect τ as the more conventional meta-analysis estimand for evidence aggregation. τ could also be of interest, because it would be the decision-maker’s prediction for the intervention’s effect in some context $k \notin \{1, 2\}$ that has not been studied:

$$\mathbb{E}[\tau_k \mid \hat{\tau}] = \mathbb{E}[\tau \mid \hat{\tau}].$$

Since the estimand can now be either a function of the effects in the study contexts or a parameter of the superpopulation model, we introduce θ as generic notation for the decision-maker’s estimand.

The process that generates study estimates and the model of harmonization remain the same, but the decision-maker now has prior beliefs about design artifact δ^j for $j \in \{1, 2\}$, which are given by $\delta^j \sim \mathcal{N}(d, v_\delta)$, where $d \in \mathbb{R}$ denotes the mean and $v_\delta > 0$ the variance. Since, the decision-maker’s prior beliefs are identical across designs, she has again no a priori reason to prefer one design over the other. This assumption resembles our notion of ambiguity-equivalence in the minimax regret framework.²

We consider two utility functions.³ First, assume the decision-maker, as in our main framework, derives utility from a binary choice $a \in \{0, 1\}$:

$$u(a; \theta) = \begin{cases} \mathbb{E}[\theta] & \text{if } a = 0 \\ \theta & \text{if } a = 1, \end{cases} \quad (29)$$

where $\mathbb{E}[\theta]$ denotes the decision-maker’s mean prior belief about θ . As before, the decision-maker is motivated to learn since she would like to chose $a = 1$ if and only if θ is large enough. Setting

²As before, the decision-maker knows the variance σ^2 of the statistical noise ϵ_i . Note that the decision-maker can estimate this variance from the data, since this variance is not impacted by the research design.

³Since harmonization (diversity) leads to a mean-preserving spread of the posterior distribution for the evidence aggregation (external validity) estimand(s), our results will generalize to any decision problem in which the decision-maker’s utility depends on the true value of the respective estimand (Blackwell, 1953).

the utility of $a = 0$ equal to the decision-maker's mean prior belief simplifies the analysis, because it makes the decision-maker indifferent between choosing $a = 1$ and $a = 0$ absent new information. Hence, new information about θ always impacts the decision-maker's choice. If the utility of $a = 0$ is not tied to the decision-maker's prior belief, this prior belief may dictate the decision-maker's choice even in the presence of new information.

Second, we consider a decision-maker who chooses to report an estimate $e \in \mathbb{R}$ of θ in order to minimize the following squared-error loss function:

$$u(e; \theta) = -(\theta - e)^2. \quad (30)$$

Our decision-maker chooses whether to harmonize, h , and makes her choice of, respectively, a or e , to maximize her expected utility.

Regardless of her utility function, the decision-maker ends up minimizing the variance $v_{post}(h) = \text{Var}(\theta \mid \hat{\tau})$ of her posterior belief about the estimand θ . $v_{post}(h)$ depends on the decision-maker's harmonization choice h , because harmonization changes the distribution of the estimate vector $\hat{\tau}$. As standard in normal-normal learning problems with known variance, $v_{post}(h)$ does not depend on the realization of $\hat{\tau}$. Hence, $v_{post}(h)$ is also the decision-maker's *expected* posterior variance at the point in time when she makes her harmonization choice, i.e., prior to the realization of $\hat{\tau}$.

Lemma A1. *For a given estimand θ , the decision-maker prefers to harmonize if and only if $v_{post}(1) < v_{post}(0)$.*

The first two panels of Figure A4 illustrate our first result. Like in our minimax regret framework, a decision-maker who seeks to aggregate evidence prefers design diversity to harmonization, irrespective of whether the estimand is the average effect across the study contexts or the superpopulation average effect.

Proposition A1. *For $\theta = \tau$ and $\theta = \frac{\tau_1 + \tau_2}{2}$, $v_{post}(1) > v_{post}(0)$ and hence $h^* = 0$, i.e., the decision-maker prefers research design diversity.*

The intuition is similar but not identical to that in the minimax regret framework. Here, design harmonization induces a correlation across the study estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ that stems from a source other than the cross-context correlation in treatment effects τ_i . By using different designs across

studies, the decision-maker can avoid this correlation. Intuitively, two independent signals about the estimand contain more information than correlated ones.

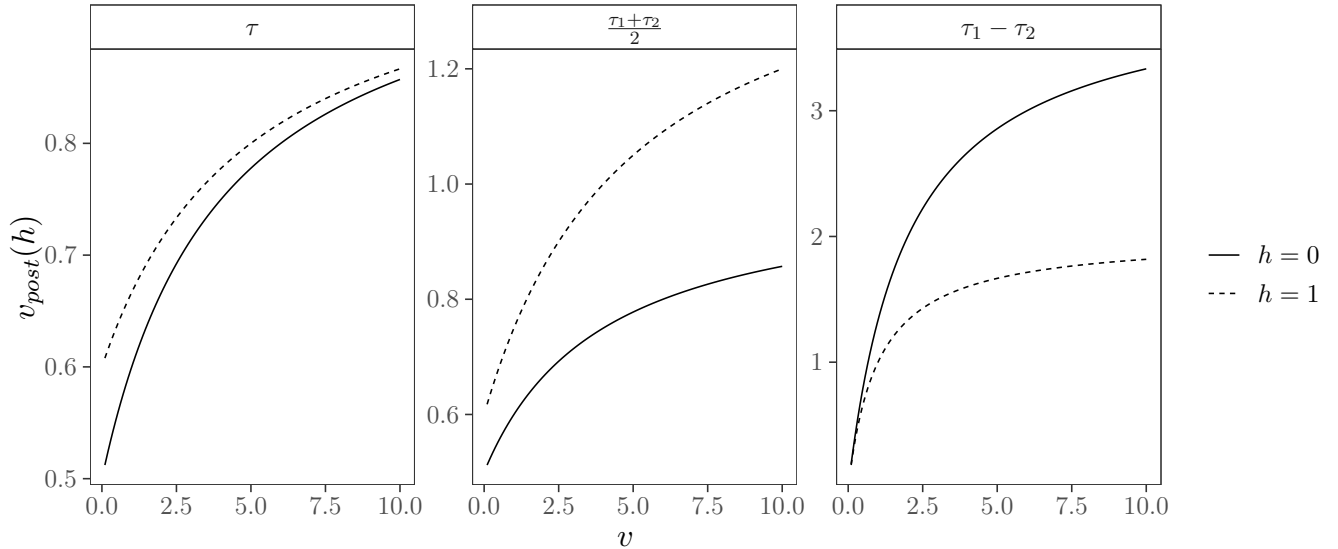


Figure A4: The variance of the decision-maker's posterior beliefs as a function of her estimand, harmonization choice and the cross-context variance in treatment effects

Plotted for $v_\delta = 1$, $\sigma^2 = 1$, $v_\tau = 1$.

The third panel in Figure A4 shows that, as in our minimax regret framework, the decision-maker prefers to harmonize if her estimand is the cross-context difference in effects. The rationale is the same. Harmonization holds design artifacts constant across context, thereby allowing the decision-maker to isolate cross-context differences in treatment effects.

Proposition A2. For $\theta = \tau_1 - \tau_2$, $v_{post}(0) > v_{post}(1)$ and hence $h^* = 1$, i.e., the decision-maker prefers research design harmonization.

F Imperfectly correlated artifacts under harmonization in the Bayesian framework

Suppose in our Bayesian framework that research design j induces design artifact δ_i^j when used in context i . Moreover, assume that the decision-maker's prior beliefs are that δ_1^j and $\delta_2^{j'}$ are independent if $j \neq j'$, i.e., as before, the decision-maker's prior beliefs under design diversity are

$$\begin{bmatrix} \delta_1^1 \\ \delta_2^2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} d \\ d \end{bmatrix}, \begin{bmatrix} v_\delta & 0 \\ 0 & v_\delta \end{bmatrix} \right).$$

However, when $j = j'$, the decision-maker expects the two design artifacts to have covariance $\kappa_\delta \in \mathbb{R}$ with $0 \leq \kappa_\delta \leq v_\delta$. Hence, under harmonization, design artifacts are distributed as follows:

$$\begin{bmatrix} \delta_1^1 \\ \delta_2^2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} d \\ d \end{bmatrix}, \begin{bmatrix} v_\delta & \kappa_\delta \\ \kappa_\delta & v_\delta \end{bmatrix} \right).$$

If $\kappa_\delta = v_\delta$, the decision-maker expects a given research design to introduce perfectly correlated artifacts across contexts as in our main framework. Harmonization leads to imperfectly correlated design artifacts as long as $0 < \kappa_\delta < v_\delta$. For $\kappa_\delta = 0$, design artifacts are completely independent across contexts even under harmonization.

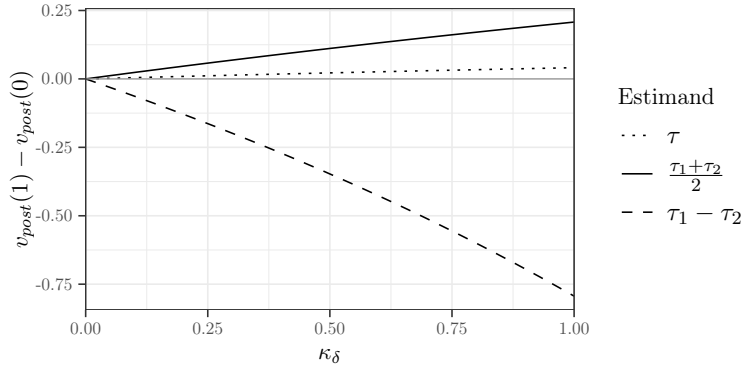


Figure A5: The difference between the posterior variance under design harmonization and diversity as a function of the cross-context covariance in design artifacts under harmonization

$v_{post}(1)$ is the posterior variance under harmonization, $v_{post}(0)$ is the posterior variance under diversity. The decision-maker seeks to minimize the posterior variance. $v_\delta = 1$, $\sigma^2 = 1$, $v_\tau = 1$, $v = 2.5$.

As a result of these changes, the covariance of the two study estimates $\hat{\tau}_1^1$ and $\hat{\tau}_2^2$ under harmonization is now given by $v_\tau + \kappa_\delta$ instead of $v_\tau + v_\delta$. Similarly, the variance of the difference in estimates under harmonization $\hat{\tau}_1^1 - \hat{\tau}_2^2$ now becomes $2(v + \sigma^2 + v_\delta - \kappa_\delta)$ instead of $2(v + \sigma^2)$. The decision-maker's posterior variance under diversity remains identical regardless of the estimand, while the posterior variance of the estimand under harmonization changes. Using the same approach as that in the proofs of propositions A1 and A2, we find the posterior variance under harmonization

takes the following form depending on the estimand:

$$v_{post}(1) = \begin{cases} \frac{(\kappa_\delta + \sigma^2 + v + v_\delta)v_\tau}{\kappa + \sigma^2 + v + v_\delta + 2v_\tau} & \text{if } \theta = \tau, \\ \frac{(\kappa_\delta + \sigma^2 + v_\delta)(v + 2v_\tau)}{2(\kappa_\delta + \sigma^2 + v + v_\delta + 2v_\tau)} & \text{if } \theta = \frac{\tau_1 + \tau_2}{2}, \\ \frac{2v(-\kappa_\delta + \sigma^2 + v_\delta)}{-\kappa_\delta + \sigma^2 + v + v_\delta} & \text{if } \theta = \tau_1 - \tau_2. \end{cases} \quad (31)$$

Using the expressions for $v_{post}(0)$ in equations 24 and 25 in the proof of proposition A1, it is easy to show that for both $\theta = \tau$ and $\theta = \frac{\tau_1 + \tau_2}{2}$, we have $v_{post}(1) \geq v_{post}(0)$ and that this inequality is strict if and only if $\kappa_\delta > 0$. In other words, for the case of evidence aggregation, the decision-maker strictly prefers design diversity as long as there is some covariance in design artifacts of a given design across contexts. Otherwise, she is indifferent. It is also the case that $\frac{\partial(v_{post}(1) - v_{post}(0))}{\partial\kappa_\delta} > 0$, i.e., the difference in the posterior variance under harmonization and diversity and hence the decision-maker's relative preference for design diversity increases the greater the cross-context covariance in design artifacts. Results for the case of external validity assessments are equivalent: Using the expression for $v_{post}(0)$ in equation 26 in the proof of proposition A2, we find that for $\theta = \tau_1 - \tau_2$, $v_{post}(0) \geq v_{post}(1)$ and that this inequality is strict if and only if $\kappa_\delta > 0$. Moreover, $\frac{\partial(v_{post}(1) - v_{post}(0))}{\partial\kappa_\delta} < 0$, i.e., the decision-maker's relative preference for harmonization becomes stronger the greater the cross-context dependence in artifacts of a given design. Figure A5 illustrates this result.

G Research design as both bias and sampling variance

What happens if research design j introduces both design artifact δ^j and variance $\sigma^{(j)2}$? We maintain the same assumptions about ambiguity and ambiguity equivalence about these parameters that we have made throughout and consider a decision-maker who faces the following decision-problem:

$$\min_h R(h), \text{ where } R(h) := \max_{\tau, \delta(h), \sigma^2(h)} r(\tau, \delta(h), \sigma^2(h)). \quad (32)$$

In the presence of design artifacts and design-specific variance, there are more cases to consider because the decision-maker's maximum regret can increase or decrease with the variance of the estimate distribution. Inspecting equation 2 in the main text, we see that whenever the numerator of the fraction inside the CDF of the standard normal distribution is negative, regret increases in the variance of the estimate distribution. Conversely, whenever the numerator of the fraction inside the

CDF of the standard normal distribution is positive, regret decreases with this variance. The latter case arises whenever the estimates are severely biased in the direction of the wrong decision. For example, if $f(\boldsymbol{\tau})$ is positive, such that the optimal decision is $a^* = 1$, and $f(\boldsymbol{\delta}(h))$ is so negative that $-f(\boldsymbol{\tau}) - f(\boldsymbol{\delta}(h)) > 0$, then the decision-maker prefers a more variable design to a design that produces estimates that are tightly concentrated around a negative mean. It follows that if the numerator of the fraction inside the CDF of the standard normal distribution is negative, the worst case regret arises whenever the variance takes its largest possible value. Conversely, whenever the numerator of the fraction inside the CDF of the standard normal distribution is positive, the worst-case obtains when the variance takes its smallest possible value.

Let's first consider the case of evidence aggregation, i.e., $f(\boldsymbol{\tau}) = \frac{\tau_1 + \tau_2}{2}$. Given the logic above, the decision-maker's maximum regret under harmonization is given by:

$$R(1) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 1) = \begin{cases} \frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{-\frac{\tau_1 + \tau_2}{2} - \underline{\delta}}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} > 0 \text{ and } -\frac{\tau_1 + \tau_2}{2} - \underline{\delta} \leq 0 \\ \frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{-\frac{\tau_1 + \tau_2}{2} - \underline{\delta}}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} > 0 \text{ and } -\frac{\tau_1 + \tau_2}{2} - \underline{\delta} > 0 \\ -\frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{\frac{\tau_1 + \tau_2}{2} + \bar{\delta}}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} \leq 0 \text{ and } \frac{\tau_1 + \tau_2}{2} + \bar{\delta} \leq 0 \\ -\frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{\frac{\tau_1 + \tau_2}{2} + \bar{\delta}}{\sigma/\sqrt{2}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} \leq 0 \text{ and } \frac{\tau_1 + \tau_2}{2} + \bar{\delta} > 0. \end{cases}$$

The decision-maker's maximum regret under diversity is

$$R(0) = \max_{\boldsymbol{\tau}} \tilde{r}(\boldsymbol{\tau}, 0) = \begin{cases} \frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{-\frac{\tau_1 + \tau_2}{2} - \underline{\delta} - \frac{\Delta}{2}}{\sqrt{\sigma^2/2 - \Delta\sigma/4}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} > 0 \text{ and } -\frac{\tau_1 + \tau_2}{2} - \underline{\delta} \leq 0 \\ \frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{-\frac{\tau_1 + \tau_2}{2} - \underline{\delta} - \frac{\Delta}{2}}{\sqrt{\sigma^2/2 + \Delta\sigma/4}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} > 0 \text{ and } -\frac{\tau_1 + \tau_2}{2} - \underline{\delta} > 0 \\ -\frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{\frac{\tau_1 + \tau_2}{2} + \bar{\delta} - \frac{\Delta}{2}}{\sqrt{\sigma^2/2 - \Delta\sigma/4}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} \leq 0 \text{ and } \frac{\tau_1 + \tau_2}{2} + \bar{\delta} \leq 0 \\ -\frac{\tau_1 + \tau_2}{2} \Phi\left(\frac{\frac{\tau_1 + \tau_2}{2} + \bar{\delta} - \frac{\Delta}{2}}{\sqrt{\sigma^2/2 + \Delta\sigma/4}}\right) & \text{if } \frac{\tau_1 + \tau_2}{2} \leq 0 \text{ and } \frac{\tau_1 + \tau_2}{2} + \bar{\delta} > 0. \end{cases}$$

It is easy to see that $R(1) > R(0)$, i.e., allowing for ambiguity over design-specific variances in addition to artifacts does not alter our conclusion that the decision-maker prefers design diversity for the purposes of evidence aggregation.

For $f(\boldsymbol{\tau}) = \tau_1 - \tau_2$, the decision-maker's regret under harmonization is given by

$$R(1) = \max_{\tau} \tilde{r}(\tau, 1) = \begin{cases} (\tau_1 - \tau_2) \Phi \left(\frac{-(\tau_1 - \tau_2)}{\sqrt{2\sigma}} \right) & \text{if } \tau_1 - \tau_2 > 0 \\ (\tau_1 - \tau_2) \Phi \left(\frac{\tau_1 - \tau_2}{\sqrt{2\sigma}} \right) & \text{if } \tau_1 - \tau_2 \leq 0, \end{cases}$$

and the decision-maker's regret under research design diversity is

$$R(0) = \max_{\tau} \tilde{r}(\tau, 0) = \begin{cases} (\tau_1 - \tau_2) \Phi \left(\frac{-(\tau_1 - \tau_2) - (\underline{\delta} - \bar{\delta})}{\sqrt{2\sigma^2 - \Delta\sigma}} \right) & \text{if } \tau_1 - \tau_2 > 0 \text{ and } -(\tau_1 - \tau_2) - (\underline{\delta} - \bar{\delta}) \leq 0 \\ (\tau_1 - \tau_2) \Phi \left(\frac{-(\tau_1 - \tau_2) - (\underline{\delta} - \bar{\delta})}{\sqrt{2\sigma^2 + \Delta\sigma}} \right) & \text{if } \tau_1 - \tau_2 > 0 \text{ and } -(\tau_1 - \tau_2) - (\underline{\delta} - \bar{\delta}) > 0 \\ (\tau_1 - \tau_2) \Phi \left(\frac{\tau_1 - \tau_2 + \bar{\delta} - \underline{\delta}}{\sqrt{2\sigma^2 - \Delta\sigma}} \right) & \text{if } \tau_1 - \tau_2 \leq 0 \text{ and } \tau_1 - \tau_2 + \bar{\delta} - \underline{\delta} \leq 0 \\ (\tau_1 - \tau_2) \Phi \left(\frac{\tau_1 - \tau_2 + \bar{\delta} - \underline{\delta}}{\sqrt{2\sigma^2 + \Delta\sigma}} \right) & \text{if } \tau_1 - \tau_2 \leq 0 \text{ and } \tau_1 - \tau_2 + \bar{\delta} - \underline{\delta} > 0. \end{cases}$$

Here, the relationship between $R(1)$ and $R(0)$ is ambiguous. On the one hand, $R(0)$ may exceed $R(1)$, because under research design diversity, the location of the study estimate distribution reflects the difference in design artifacts in addition to the cross-context difference in treatment effects. On the other hand, $R(0)$ may be smaller, since the smallest possible variance of the estimate distribution is larger and the largest possible variance of the estimate distribution is smaller under design diversity than under research design harmonization. Whether the decision-maker prefers to harmonize or to diversify for the purposes of external validity assessments depends on the magnitude of these two effect.

H Within-study design diversity

Suppose our decision-maker runs an experiment in a single context i with n_i units. In what follows, we omit the subscript i for brevity, since we only consider a single context in this section. We use l to index units with $l = 1, \dots, n$. Let $z_l \in \{0, 1\}$ denote the treatment status of unit l . Invoking the stable unit treatment value assumption (SUTVA), we write unit l 's potential outcomes as $y_l(z)$. The decision-maker's inferential target is the average treatment effect (ATE) across the n units in the experiment:

$$\tau = \frac{1}{n} \sum_{l=1}^n y_l(1) - y_l(0).$$

The decision-maker uses complete random assignment to assign m units to treatment and $n - m$ units to control. The decision-maker can choose between two outcome measures $j \in \{1, 2\}$. Outcome measures introduce systematic measurement error. In particular, outcome measure j adds a constant $\nu^j \in \mathbb{R}$ to the treated potential outcome for every subject, e.g., because the treatment causes experimenter demand effects.⁴ Hence, given research design j , subject l 's *measurable* potential outcomes $\tilde{y}_l(z_l)$ are

$$\begin{aligned}\tilde{y}_l^j(1) &= y_l(1) + \nu^j \\ \tilde{y}_l^j(0) &= y_l(0).\end{aligned}$$

The decision-maker faces ambiguity over measurement errors. As before, we assume that $\nu^j \in [\underline{\nu}, \bar{\nu}]$ for $j \in \{1, 2\}$ with $\underline{\nu}, \bar{\nu} \in \mathbb{R}$ and $\underline{\nu} < \bar{\nu}$ as well as $|\nu^1 - \nu^2| \geq \Delta_\nu > 0$.

The decision-maker again makes a choice $h \in \{0, 1\}$ of whether to harmonize. Within-study harmonization here entails using the same outcome measure – w.l.o.g. measure 1 – for all n units in the experiment. A decision-maker who induces within-study diversity randomly assigns each unit to one of the two outcome measures. For simplicity, we assume the decision-maker assigns units to outcome measure 1 with exogenously given probability $\gamma \in (0, 1)$ and to outcome measure 2 with complementary probability $1 - \gamma$. The assumption here is that the decision-maker cannot use both outcome measures for all units, e.g., because of a budget constraint.

Once she has run the experiment, the decision-maker uses the difference-in-means estimator to generate an estimate $\hat{\tau}^j$ of τ using the observed outcomes:

$$\hat{\tau}^j = \frac{1}{m} \sum_{l=1}^m \tilde{y}_l^j(1) - \frac{1}{n-m} \sum_{l=m+1}^n y_l(0),$$

where we assume that the units in our subject pool have been ordered by treatment status with the first m units being the treated ones. As before, the decision-maker makes a binary choice $a \in \{0, 1\}$

⁴This is a simplified version of the more complicated measurement error model discussed in appendix section C.

which generates a payoff that depends on the estimand, here τ :

$$u(a, \tau) = \begin{cases} 0 & \text{if } a = 0 \\ \tau & \text{if } a = 1. \end{cases} \quad (33)$$

Moreover, the decision-maker again uses a decision-rule where she chooses $a = 1$ if and only if $\hat{\tau}^j > 0$. The decision-maker's regret is defined as before:

$$r(\tau, h) = u(a^*; \tau) - \mathbb{E}[u(a; \tau)],$$

where the expectation in the second part of the expression is now taken over the distribution of the difference-in-means estimator induced by the random assignment of treatment across repeated experiments. As before, the decision-maker chooses whether to harmonize in order to minimize her maximal regret:

$$\min_h R(h), \text{ where } R(h) := \max_{\tau, \boldsymbol{\nu}(h)} r(\tau, \boldsymbol{\nu}(h)).$$

$\boldsymbol{\nu}(h)$ is an n -dimensional vector where the l th element refers to the measurement error of the l th unit in the experiment.

The choice of whether to harmonize affects both the expectation and the variance of the sampling distribution of the difference-in-means estimator. The expectation of the sampling distribution is given by

$$\mathbb{E}[\hat{\tau}] = \begin{cases} \tau + \gamma\nu^1 + (1 - \gamma)\nu^2 & \text{if } h = 0 \\ \tau + \nu^1 & \text{if } h = 1. \end{cases}$$

Intuitively, the location of the sampling distribution reflects both design artifacts weighted by the share of units assigned to each outcome measure under design diversity and only the artifact of design 1 under design harmonization. The variance of the difference-in-means estimator, following [Neyman](#)

(1923), is given by

$$\sigma^2(h) = \frac{S_1^j}{m} + \frac{S_0}{n-m} - \frac{S_{10}^j}{n},$$

where

$$\begin{aligned} S_1^j &= \frac{1}{n-1} \sum_{l=1}^n (\tilde{y}_l^j(1) - \bar{\tilde{y}}_1^j)^2 \\ S_0 &= \frac{1}{n-1} \sum_{l=1}^n (y_l(0) - \bar{y}_0)^2 \\ S_{10}^j &= \frac{1}{n-1} \sum_{l=1}^n (\tilde{\tau}_l^j - \bar{\tau}^j)^2. \end{aligned}$$

Here, $\bar{\tilde{y}}_1^j$ is the average of the *measurable* treated potential outcomes across the N experimental units and \bar{y}_0 the average of the control potential outcomes, since we have assumed that control potential outcomes are not subject to measurement error. $\tilde{\tau}_l^j$ is the difference between individual l 's measurable treated potential outcome $\tilde{y}_l^j(1)$ and her control potential outcome $y_l(0)$, and $\bar{\tau}^j$ the average of this difference across the n experimental units.

How does this variance depend on the decision-maker's harmonization choice h ? Note first that given our very simple model of measurement error, the variance $\sigma^2(1)$ under harmonization equals the variance that would obtain if the decision-maker could directly observe $y_l(1)$ and not just $\tilde{y}_l^j(1)$ for each unit. The reason is that, in this case, measurable treated potential outcomes $\tilde{y}_l^j(1)$ simply equal the sum of each unit's treated potential outcome $y_l(1)$ and the constant ν^1 . Adding a constant has no effect on the variance of the treated potential outcomes and hence neither on S_1^j and S_{10}^j . Under research design diversity, however, the variance of the measurable treated potential outcomes increases. The reason is that measurable treated potential outcomes $\tilde{y}_l^j(1)$ are now the sum of each unit's treated potential outcome $y_l(1)$ and a *random variable* that takes the value ν^1 with probability γ and the value ν^2 with probability $1 - \gamma$. Since the decision-maker randomly assigns units to outcome measures, the design artifact is independent from treated potential outcomes. Hence, the variance of the measurable treated potential outcomes $\tilde{y}_l^j(1)$ is simply the sum of the variance of the treated potential outcomes $y_l(1)$ and $\gamma(1 - \gamma)(\nu_1 - \nu_2)^2$. S_1^j is thus greater under research diversity than under harmonization. Random assignment of outcome measures to units also implies that the

covariance of measurable treated and untreated potential outcomes and hence S_{10}^j remains unaffected by the harmonization choice. It follows that, for any given constellation of potential outcomes, the variance of the difference-in-means estimator is greater under research design diversity than under research design harmonization: $\sigma^2(0) > \sigma^2(1)$.

The CLT implies that the standardized difference-in-means converges to a standard normal distribution (Li and Ding, 2017). Using this fact, we can approximate the decision-maker's regret under design harmonization by

$$r(\tau, 1) = \begin{cases} \tau \Phi\left(\frac{-\tau - \nu^1}{\sigma(1)}\right) & \text{if } \tau > 0 \\ -\tau \Phi\left(\frac{\tau + \nu^1}{\sigma(1)}\right) & \text{if } \tau \leq 0, \end{cases} \quad (34)$$

and under design diversity by

$$r(\tau, 0) = \begin{cases} \tau \Phi\left(\frac{-\tau - \gamma\nu^1 - (1-\gamma)\nu^2}{\sigma(0)}\right) & \text{if } \tau > 0 \\ -\tau \Phi\left(\frac{\tau + \gamma\nu^1 + (1-\gamma)\nu^2}{\sigma(0)}\right) & \text{if } \tau \leq 0. \end{cases} \quad (35)$$

Our main take-away from expressions 34 and 35 is that the choice between harmonization and diversity within a study can involve a bias-variance trade-off. The largest possible shift of the estimate distribution under research-design harmonization is $\bar{\nu}$ and the smallest is $\underline{\nu}$. Under research-design diversity, the largest possible bias is $\bar{\nu} - \min\{\gamma, 1 - \gamma\}\Delta_\nu$ and the smallest is $\underline{\nu} + \min\{\gamma, 1 - \gamma\}\Delta_\nu$. Hence, the largest possible bias is larger and the smallest possible bias is smaller under research design harmonization than under research design diversity. Hence, if we focus only on shifts in the location of the estimate distribution, research design diversity is preferred. The logic is the same as the one in the case of evidence aggregation in our main model: Research design diversity guards against the worst-case scenario of using a single highly biased outcome measure for all units in the experiment. However, we also know from the analysis above that $\sigma(0) > \sigma(1)$ and as long as the worst-case biases are not too big in absolute value, the decision-maker prefers estimates that are less variable. Hence, a bias-variance trade-off may arise where design diversity minimizes bias but harmonization maximizes precision. Whether this trade-off arises and which concern dominates depends on the particular constellation of potential outcomes and the size of the worst-case biases.

However, the decision-maker can eliminate the impact of her harmonization decision on the vari-

ance of her estimates using a design-based strategy. Suppose the decision-maker first randomly assigns units to one of the two outcome measures and then uses *block-random* assignment to assign units to treatment where the blocks are formed based on the outcome measures to which units were assigned. With block-random assignment, the variance of the difference-in-means estimator is given by

$$\sigma_{blocked}^2 = \sum_{j=1}^2 \frac{n_j^2}{n^2} \text{var}(\hat{\tau}_j^j),$$

where $\text{var}(\hat{\tau}_j^j)$ is the variance of the estimates from the block with units assigned to measure j . Note that, since units in block j are all assigned to the same outcome measure which adds the same constant ν_j to their treated potential outcomes, estimates under research design diversity with this blocking structure are again no more variable than estimates under harmonization or in a world where the researcher can directly observe treated potential outcomes of units in the treatment group without measurement error. As a result, the choice between within-study harmonization and diversity hinges on bias-concerns only and on this dimension, design diversity minimizes the decision-maker’s maximum regret.

References

- Berger, Roger L and Jason C Hsu. 1996. “Bioequivalence trials, intersection-union tests and equivalence confidence sets.” *Statistical Science* 11(4):283–319.
- Blackwell, David. 1953. “Equivalent comparisons of experiments.” *The annals of mathematical statistics* pp. 265–272.
- DeGroot, Morris H. 2005. *Optimal Statistical Decisions*. John Wiley & Sons.
- Egami, Naoki and Erin Hartman. 2023. “Elements of external validity: Framework, design, and analysis.” *American Political Science Review* 117(3):1070–1088.
- Greene, William H. 2011. “Econometric analysis 7th edition.” *International edition, New Jersey: Prentice Hall* .
- Li, Xinran and Peng Ding. 2017. “General forms of finite population central limit theorems with applications to causal inference.” *Journal of the American Statistical Association* 112(520):1759–1769.

Lord, Frederic M and Melvin R Novick. 2008. *Statistical theories of mental test scores*. IAP.

Neyman, Jersey. 1923. “Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes.” *Roczniki Nauk Rolniczych* 10(1):1–51.

Schlaifer, Robert and Howard Raiffa. 1961. *Applied statistical decision theory*.