# SYNTHESIS OF ALTERNATE EVALUATION MEASURES OF PUBLIC EDUCATION CAMPAIGNS

**David Judkins and Paul Zador, Westat**
**David Judkins, Westat, 1650 Research Boulevard, Rockville, Maryland 20850**

**Key Words:** Propensity scoring, causal inference, dose-response, trend test, confounding

## 1. Introduction

Public education campaigns such as that described by Lu et al (2001) consist of delivering small packets of information in a largely uncontrolled manner over a large area for an extended period of time. The purpose of such campaigns is to change the knowledge, thinking, and behavior of the population. Such campaigns pose some special problems for evaluation, particularly when the campaign is national in scope and evaluation efforts are commenced after the campaign launch. A national scope precludes the possibility of using control areas in evaluation, and a post-launch evaluation commencement precludes the use of before and after designs. The only remaining options are to study dose-response relationships and to monitor for continuing temporal change as the campaign continues.

The evaluation method adopted by Hornik et al (2001) for the same evaluation project referenced by Lu and coauthors is a synthesis of analysis of temporal change and of the dose-response relationship between campaign exposure and targeted outcomes. Referring to the typology of designs for quasi-experimentation laid out by Cook and Campbell (1979), the temporal change analysis is a variant of the one-group pretest-posttest design, and the dose-response analysis is a variant of the posttest-only design with nonequivalent groups, both of which Cook and Campbell classified as facing serious inferential threats. Given that the inferential threats are different for the two designs, Hornik et al adopted the stance that the campaign would be judged successful only if both types of analysis found evidence of consistent effects. We demonstrate that this approach is rather conservative unless the sizes of the tests used in each analysis are adjusted. Prior to proving this assertion, the well known weaknesses of each method are briefly reviewed as motivation for the synthetic analysis.

## 2. Interpretation of Temporal Trends

Temporal succession of causes and effects is a fundamental aspect of concepts of causation. Holland (1986) touches upon this in his brief review of the causal theories of Hume, Mill, Suppes, and Hill. If an intervention with persistent effects is running at a fairly even intensity for an extended period, then, in the absence of other forces on society, one would expect a time series of the population to show a trend in the desired direction. Obviously, however, the targets of a public education campaign are buffeted by many other forces in their daily lives.

For evaluation of social interventions, the interrupted time series is a plausible tool. With such an approach, a series of observations are made during the course of which the intervention is turned on and off to see if Hume's condition for constant conjunction is met. Since the uncontrolled forces would be presumed to vary over time, if the outcome timeseries always shifts shortly after restarting or interrupting the time series, then the inability to isolate the system from other forces would not be particularly worrisome. However, public education campaigns, like many other social interventions, are not easy to interrupt. These interventions require systems development, trained personnel and management. Costs for dismissing and then rehiring and retraining intervention workers around the interruptions may be high. Proponents of the intervention may lobby against interruptions. Some interventions may also have persistent and lagging effects, so the interruptions may need to be fairly long, depending on societal relaxation time.

So the typical evaluation of any social intervention will involve tracking of outcomes over time as the intervention proceeds according to its own dynamics, uninfluenced by the evaluation. Other forces will also continue according to their own dynamics. Supporters of the intervention see the intervention as responsible for any positive trends while extrinsic forces are blamed for negative trends. Opponents are likely to reverse these rules.

## 3. Interpretability of Dose-Response Relationships

From data collected at a single point in time (meaning, in actuality, a short period of time), any effect of the intervention should manifest itself in a systematic difference between treated and untreated cases, or, in the case of ordinal-measured intervention intensity, a monotone relationship between exposure and response. The latter refers back to Hill's concept of a biological gradient (Holland, 1986). In the study discussed by Lu et al and by Hornik et al, doses were not randomly assigned and were only measured retrospectively. The threats to inference here include the failure to measure common causes (which lead to "self-selection bias"), errors in the measurement of exposure (the dose), errors in the measurement of the outcome variable (the response), and errors in temporal sequencing of exposure and outcomes. These threats are mitigated primarily through the development of good measurement instruments and protocols. Of course, these are only limited by the imagination and skills of the instrument and protocol designers. However, as many prior

authors have discussed, the existence of unmeasured common causes can never be disproven from data from a nonrandomized study. Most famously, this was the basis of R. A. Fisher's steadfast rejection of studies that claimed a causal link between tobacco smoking and cancer (Box, 1978, pp 472-476).

Nor can the nonexistence of measurement errors be proven. If exposure measurement errors are independent of all other variables, then their only effect is to bias the study toward finding no effect. However, if exposure measurement errors are correlated with outcomes, it is possible to have strong biases toward findings of effects or counter-effects (where the opposite of the desired effect is observed). This is of particular concern when the outcomes are cognitive rather than behavioral and where exposure is measured retrospectively. If cognitive dissonance degrades memory (Festinger, 1957; and Wickland and Brehm, 1976), then the effect of the campaign will be exaggerated in that people who already agreed with the messages report higher levels of exposure and those who already disagreed report lower levels of exposure. These threats could be partially addressed by collecting retrospective data on prior states of the outcome variables (i.e., "how did you feel before you saw the commercial?"), but this is a messy process with cognitive outcomes. If reports of past levels of a cognitive outcome are biased toward harmony with current levels, then such a corrective approach would be defeated. Perhaps this difficulty can be surmounted by identifying and measuring common causes for past cognitions and measurement errors, but uncertainty will remain.

Finally, there are questions about event sequencing. Since both exposure and cognitive outcomes develop over time but are measured only once in a retrospective study, it is possible that changes in cognitive variables lead to changes in true exposure. An obvious example would be an advertising campaign that encouraged people to turn off their TV sets. For those subjects where the campaign was successful, there would be no recent exposure to report.

Thus, if a significant dose-response relationship is found, supporters of the intervention will tend to believe that all relevant common causes were measured, that there were no important correlations between measurement errors and other variables, and that time sequence information is correct. Opponents of the intervention are more likely to invent plausible unmeasured common causes, correlated measurement errors, and time-sequence errors, or to simply insist that nothing is proven without randomization.

## 4. Synthesis

Given the weakness of both methods, the procedure of performing both types of analysis has some appeal. Since the threats to inference are completely different for the two methods, if the two methods agree, then one might feel somehow more confident in one's conclusions. Along these lines, Rosenbaum (2001) advocated replicating scientific studies in ways that disrupt features so as to induce different biases. We suspect that this greater confidence has to do with personal subjective probabilities. Somehow it seems less likely that both sets of threats are present even though no formal assessment of that likelihood is possible and there is certainly nothing to rule out the possibility that both types of inferential threats are present. There are echoes here of Campbell (1963) that experiments probe theory, but do not prove theory. The problem with requiring significant results from both of the two methods is that this synthesis dramatically changes the balance between formal type I and type II errors unless the standard of significance is relaxed for one or both of the tests. How this balance changes will depend on the nature of the stochastic process induced by the campaign on the population. In Section 5, we present a reasonable model for this process and then demonstrate that, for this model at least, the tests for temporal change and a dose-response relationship are asymptotically independent of each other.

If two independent tests of size $\alpha$ of the same set of hypotheses are carried out, then it is obvious that the size of the synthesized test is $\alpha^2$. If the power of the change test at a specified point in the alternative hypothesis universe is $1-\beta_1$ and the power of the dose-response test at the same point is $1-\beta_2$, then it is also obvious that the power of the synthesized test at that point will be $(1-\beta_1)(1-\beta_2)$. If standard testing procedures are followed with $\alpha = 0.05$ and the sample sizes are large enough that $\beta_1 = \beta_2 = 0.2$, such a double hurdle will result in a synthesized test of size of 0.0025 and power of 0.64. For some applications where it is desired to give a very strong benefit of the doubt to the null hypothesis of no effect, this may be a sensible testing procedure. However, it is a more stringent standard of evidence than is usually set in biopharmaceutical research, epidemiological research, psychometric research, and indeed, most types of research. While this may be appropriate in the evaluation of some interventions, this adoption of a more stringent criterion should be made consciously. We suggest that if this double-hurdle approach is to be used, then the size of each test should be adjusted so that the size of the synthesized test is $\alpha$ and a more appropriate balance between type I and type II errors is thereby achieved. One obvious choice would be to set the size of each at $\sqrt{\alpha}$, but any factorization $\alpha_1\alpha_2 = \alpha$ could be used. For example, if the inferential threats to the trend analysis were viewed as more serious than those to the dose-response analysis, then a factorization of (0.5)(0.1)=0.05 might be used, meaning that an effect would be concluded if the sign of the trend test was consistent with a dose-response test that would ordinarily be characterized as merely presenting some evidence.

## 5. A Model for the Effects on a Continuing Public Education Campaign

A model for the effect of an ongoing media campaign is that each successive exposure has a small incremental effect. One hopes that the effects persist (although some relaxation is to be expected with time). Assume that the confounders for exposure and the outcome of interest have been identified so that the exposed-at-random assumption is justified. This requires that all analyses control on that set of confounders through stratification or other means. To simplify notation, the model in this paper is set up within a single stratum of such a stratification and the stratum notation is suppressed. Also, time is treated as discrete. We first deal with a simple pre-post measurement in conjunction with a dose-response relationship and then extend it to multiple post-implementation measurement points without a pre-campaign measurement. With this preamble, the first model to be studied here is that

$$y_{0i} = u_0 + e_{0i}, \text{ and}$$
$$y_{1i} = u_1 + \Delta_1 d_{1i} + e_{1i}$$

where

$y_{ti}$ is the outcome variable[1] at time $t$ for subject $i$;

$n_t$ is the sample size at time $t$;

$(u_0, u_1)$ is the stochastic process that would have occurred in the absence of the campaign with mean $(\lambda_0, \lambda_1)$, and arbitrary covariance structure;

$\Delta_1$ is the effect of the campaign on an exposed subject at time 1;

$d_{1i}$ is a binary flag for exposure up between items 0 and 1 for subject $i$ with $\mathrm{E}d_{1i} = p_1$, $\mathrm{Var}d_{1i} = p_1 q_1$, where $q_1 = 1 - p_1$, and the flag is independently and identically distributed across subjects;

$(e_{0i}, e_{1i})$ is a random error vector for the $i$-th subject that is independent across subjects with constant mean (0,0), variance $\left(\sigma_0^2, \sigma_1^2\right)$, and arbitrary autocorrelation; and

$d_{1i}$, $(u_0, u_1)$ and $(e_{0i}, e_{1i})$ are all mutually independent.

In the simple case where there is a pretest and a single post-test and $\lambda_1 = \lambda_0$, an unbiased estimate of $\Delta_1$ based on temporal change is

$$\hat{\Delta}_1' = \frac{\sum y_{1i} / n_1 - \sum y_{0i} / n_0}{\sum d_{1i} / n_1},$$

and an unbiased estimate of $\Delta_1$ based on the dose-response relationship is

$$\hat{\Delta}_1'' = \frac{\sum d_{1i} y_{1i}}{\sum d_{1i}} - \frac{\sum (1 - d_{1i}) y_{1i}}{\sum (1 - d_{1i})}.$$

In this simple case, one could test for $\Delta_1 \neq 0$ by separately performing Z-tests with both $\hat{\Delta}_1'$ and $\hat{\Delta}_1''$, but if one were to conduct such a pair of tests, it is easy to demonstrate that the two statistics are uncorrelated.

**Theorem 1.** If the samples at times 0 and 1 are independent of each other, then $Cov(\hat{\Delta}_1', \hat{\Delta}_1'') = 0$.

**Proof:** Using standard conditioning arguments,

$$\begin{aligned}\mathrm{Cov}\left(\hat{\Delta}_1', \hat{\Delta}_1'',\right) = {} & \mathrm{Cov}\left[\mathrm{E}\left(\hat{\Delta}_1' \middle| u, d\right), \mathrm{E}\left(\hat{\Delta}_1'' \middle| u, d\right)\right] \\ & + \mathrm{E}\left\{\mathrm{Cov}\left[\mathrm{E}\left(\hat{\Delta}_1' \middle| u, d\right), \mathrm{E}\left(\hat{\Delta}_1'' \middle| u, d\right) \middle| u\right]\right\} \\ & + \mathrm{E}\left\{\mathrm{E}\left[\mathrm{Cov}\left(\hat{\Delta}_1', \hat{\Delta}_1'' \middle| u, d\right) \middle| u\right]\right\}.\end{aligned}$$

Each of these three terms is equal to zero. To see this for the first two terms, note that

$$\mathrm{E}\left(\hat{\Delta}_1'' \middle| u, d\right) = \mathrm{E}\left(\frac{(u_1 + \Delta_1)\sum d_{1i}}{\sum d_{1i}} - \frac{u_1 \sum (1 - d_{1i})}{\sum (1 - d_{1i})} \middle| u, d\right) = \Delta_1$$

depends on neither $u$ nor $d$, so the covariance of it with $\mathrm{E}\left(\hat{\Delta}_1' \middle| u, d\right)$ must be zero—both unconditionally (as in the first term) and when conditioned on $u$ (as in the second term).

By the independence of the two samples, the conditional covariance within the third term is equal to

$$\begin{aligned}\mathrm{Cov}\left(\hat{\Delta}_1', \hat{\Delta}_1'' \middle| u, d\right) = {} & \mathrm{Cov}\left[\frac{\sum d_{1i} e_{1i}}{\sum d_{1i}}, \frac{\sum e_{1i}}{n_1} \middle| d, u\right] \\ & - \mathrm{Cov}\left[\frac{\sum (1 - d_{1i}) e_{1i}}{\sum (1 - d_{1i})}, \frac{\sum e_{1i}}{n_1} \middle| d, u\right] \\ = {} & \frac{\sum d_{1i} \sigma_1^2}{n_1 \sum d_{1i}} - \frac{\sum (1 - d_{1i}) \sigma_1^2}{n_1 \sum (1 - d_{1i})} = 0.\end{aligned}$$

QED

In the case of study discussed by Lu et al and by Hornik et al, there was, however, no pretest, and there was a series of post measurements to capture what was theorized to be a gradual progression. So moving now to an arbitrary number of post-campaign implementation measurements, a more general model is that

$$y_{ti} = u_t + \sum_{s=1}^{t} \Delta_s d_{si} + e_{ti} \quad \text{for } t=1,\ldots \text{ and } i=1,\ldots, n_t,$$

where

$\{u_t\}_t$ is the stochastic process that would have occurred in the absence of the campaign with mean $\lambda_t$, and arbitrary covariance structure;

[1] A similar model can be developed for a binary outcome using a logit or probit transform.

$\Delta_t$ is the effect of the campaign on an exposed subject at time *t*;

$\{d_{ti}\}_t$ is a series of independent[2] binary flags for exposure[3] up through time *t* for subject *i* with $\mathrm{E}d_{ti} = p_t$, $\mathrm{Var}d_{ti} = p_t q_t$, where $q_t = 1 - p_t$, and the series are independently[4] and identically distributed across subjects;

$\{e_{ti}\}_t$ is a stochastic error process for the *i*-th subject that is independent[5] across subjects with constant mean 0, variance $\sigma_t^2$, and arbitrary autocorrelations; and

$\{d_{ti}\}_t$, $\{u_t\}_t$ and $\{e_{ti}\}_t$ are all mutually independent.[6]

Note that the expected outcome at time *t* for subjects who were exposed between times *t*-1 and *t* is

$$\varphi_t = \alpha + \lambda_t + \sum_{s=1}^{t-1} \Delta_s p_s + \Delta_t ,$$

the expected outcome for those not exposed between those times is

$$\theta_t = \alpha + \lambda_t + \sum_{s=1}^{t-1} \Delta_s p_s ,$$

and that the overall expected outcome is

$$\mu_t = \alpha + \lambda_t + \sum_{s=1}^{t} \Delta_s p_s .$$

An unbiased estimator for the mean of the recently treated subjects is

$$\hat{\varphi}_t = \frac{\sum_i d_{ti} y_{ti}}{n_t \hat{p}_t}, \text{ where } \hat{p}_t = \frac{\sum_i d_{ti}}{n_t} .$$

An unbiased estimator for the mean of the recently untreated subjects is

$$\hat{\theta}_t = \frac{\sum_i (1 - d_{ti}) y_{ti}}{n_t \hat{q}_t}, \text{ where } \hat{q}_t = \frac{\sum_i (1 - d_{ti})}{n_t} .$$

An unbiased estimator for the mean of all subjects at time *t* is

$$\hat{\mu}_t = \frac{\sum_i y_{ti}}{n_t} .$$

An unbiased estimator for the average short-term effect of treatment $\bar{\Delta} = \sum \Delta_s / t$ up through time *t* is

$$\hat{\Delta} = \frac{1}{t} \sum_{s=1}^{t} \left( \hat{\varphi}_s - \hat{\theta}_s \right) .$$

An unbiased estimator of the average change per unit time from time 1 to time *t* is

$$\hat{\Psi} = \frac{(\hat{\mu}_t - \hat{\mu}_1)}{t - 1} .$$

Note that if the underlying stochastic process is stationary so that $\lambda_t \equiv \lambda$, then

$$\hat{\Omega} = \sum_{s=2}^{t} \frac{(\hat{\mu}_s - \hat{\mu}_{s-1})}{(t-1)\, \hat{p}_s}$$

is also an unbiased estimator of $\bar{\Delta}$ so that one could synthesize the two types of analysis by using a linear combination of $\hat{\Delta}$ and $\hat{\Omega}$ to estimate $\bar{\Delta}$. However, such a statistic would be vulnerable to both types of inferential threats—those that threaten $\hat{\Delta}$ and those that threaten $\hat{\Omega}$.

The prime question for this paper is the relationship between $\hat{\Delta}$ and $\hat{\Psi}$. Due to the fact that a large value of $\bar{\Delta}$ in the absence of underlying process drift will tend to make both $\hat{\Delta}$ and $\hat{\Psi}$ large, some researchers' intuition is that the two statistics are positively correlated. It is proven in the appendix that this is false. They are, in fact, asymptotically uncorrelated, as stated in Theorem 2. If the two statistics have a joint asymptotic bivariate normal distribution, this is enough to demonstrate asymptotic independence.

**Theorem 2.** If $t = 2$ and if $n_2$ is large enough and $p_2$ is sufficiently far from 1 so that $E\left(\frac{\hat{p}_2}{\hat{q}_2}\right) = \frac{p_2}{q_2} + \mathrm{O}(1/n)$ and $E\left(\frac{1}{\hat{q}_2}\right) = \frac{1}{q_2} + \mathrm{O}(1/n)$, then $\mathrm{Cov}\left(\hat{\Delta}, \hat{\Psi}\right) = 0 + \mathrm{O}(1/n)$.

**Proof:** See appendix.

**Note:** If the sample sizes and exposure probabilities are such that the chance of a sample where everyone is either exposed or unexposed is not vanishingly small, then in practice, the stratum would be collapsed with another stratum.

---

[2] If there was any serial correlation within this series, then past exposure would become an unmeasured common cause since it would affect both exposure at time *t* and the counterfactual outcome pertinent to a stoppage of the campaign at time *t*-1. For the dose-response analysis at time *t* to be valid, it is required that the stratification be fine enough so that there are no remaining systematic sources of variability in exposure within each stratum.

[3] With discrete time, exposure at time *t* is conceptualized as exposure above some threshold between times *t-1* and *t*.

[4] We think that the independence of exposure across subjects could probably be weakened without altering the principal result of the paper.

[5] Similarly, we think that the results would generalize to a situation where there was intra-class correlation of outcomes.

[6] The assumption that $\{d_{ti}\}_t$ is independent of both $\{u_t\}_t$ and $\{e_{ti}\}_t$ is the key assumption that exposure is conditionally independent of the process that would have occurred in the absence of the campaign. As noted earlier, this is only assumed to hold within strata, the notation for which is suppressed.

## 6. Summary

For the evaluation of public education campaigns that are national in scope and start before the evaluation starts, it might be reasonable to measure both temporal trends and dose-response relationships. If both approaches are used, then at some point, it will be necessary to try to reconcile results from the two. When this is done, it is important to consider the joint properties of test statistics from the two approaches. Using a typical standard of evidence for each of the two prior to synthesis results in a nearly insuperable standard of evidence for the efficacy of the campaign. We have indicated how to adjust the sizes of the hypothesis tests associated with each approach so as to attain the desired overall test size. The details depend on the perceived inferential threats attendant upon each approach. Unless there is substantial doubt that a reasonably adequate set of confounders has been measured and there is great confidence that the time series would have been flat in the absence of the campaign, our preference would be to maintain a high standard of evidence for the dose-response analysis and largely ignore the temporal analysis. If the study were to continue for several years, one might consider eventually giving more weight to the temporal analysis. However, for early reports, the dose-response analysis seems to be the best approach.

## 7. References


Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist.* New York: John Wiley and Sons.

Campbell, D. T. (1963). From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.) *Problems in Measuring Change*. Madison, WI: U. of Wisc. Press.

Cook, T. D. and Campbell, D. T. (1979), *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Boston: Houghton Mifflin.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945-960.

Hornik, R., Maklan, D., Judkins, D., D., Cadell, Yanovitzky, I., Zador, P., Southwell, B., Mak, K., Das, B., Prado, A., Barmada, C., Jacobsohn, L., Morin, C., Steele, D., Baskin, R., and Zanutto, E. (2001). Evaluation of the National Youth Anti-Drug Media Campaign: Second Semi-Annual Report of Findings - April 2001. Rockville, Maryland: Westat.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001), Matching with Doses in an Observational Study of a Media Campaign Against Drug Abuse, *Journal of the American Statistical Association*, 96, 1,245-1,253.

Rosenbaum, P. R. (2001). Repeating effects and biases. *The American Statistician*, 55, 223-227.

Wickland, R. & Brehm, J. (1976). *Perspectives on Cognitive Dissonance*. NY: Halsted Press.


## Appendix

For simplicity, the proof is restricted to two timepoints. Assume that two nonoverlapping samples are selected at the two time points and that the population is large enough so that the finite population correction factor can be ignored. Note that we have two samples of $\{d_{ti}\}$ at time 1. These will be denoted as $\{d'_{1i}\}$ and $\{d''_{1i}\}$, respectively. Their sample means will be denoted as $\hat{p}'_1$ and $\hat{p}''_1$. Before demonstrating the main result, it will be useful to state several lemmas.

**Lemma 1.** The conditional expectation of $\hat{\Delta}$ given $\{d\}$ and $\{u\}$ is $\mathrm{E}\left(\hat{\Delta}\middle|d,u\right)=\frac{\Delta_1}{2\hat{q}_2}(Z-\hat{p}'_1)+\bar{\Delta}$,

where $Z=\frac{\sum_i d_{2i}d''_{1i}}{\sum_i d_{2i}}$.

**Lemma 2.** $\mathrm{E}\left(\hat{\Delta}\middle|u\right)=\bar{\Delta}$.

**Proof:** Since both samples are unbiased, $\mathrm{E}(Z)=p_1$ and $\mathrm{E}(\hat{p}''_1)=p_1$.

**Lemma 3.** $\mathrm{E}\left(\hat{\Psi}\middle|d,u\right)=\Delta_2\hat{p}_2+(u_2-u_1)+\ \Delta_1(\hat{p}''_1-\hat{p}'_1)$ and $\mathrm{E}\left(\hat{\Psi}\middle|u\right)=\Delta_2 p_2+(u_2-u_1)$.

**Theorem 2.** If $t=2$ and if $n_2$ is large enough and $p_2$ is sufficiently far from 1 so that $E\left(\frac{\hat{p}_2}{\hat{q}_2}\right)=\frac{p_2}{q_2}+\mathrm{O}(1/n)$ and $E\left(\frac{1}{\hat{q}_2}\right)=\frac{1}{q_2}+\mathrm{O}(1/n)$, then $\mathrm{Cov}\left(\hat{\Delta},\hat{\Psi}\right)=0+\mathrm{O}(1/n)$.

**Proof:** Using standard conditioning arguments,

$$\begin{aligned}\mathrm{Cov}\left(\hat{\Delta},\hat{\Psi}\right) &= \mathrm{Cov}\left[\mathrm{E}\left(\hat{\Delta}\middle|u\right),\mathrm{E}\left(\hat{\Psi}\middle|u\right)\right]\\ &+\mathrm{E}\left\{\mathrm{Cov}\left[\mathrm{E}\left(\hat{\Delta}\middle|u,d\right),\mathrm{E}\left(\hat{\Psi}\middle|u,d\right)\middle|u\right]\right\}\\ &+\mathrm{E}\left\{\mathrm{E}\left[\mathrm{Cov}\left(\hat{\Delta},\hat{\Psi}\middle|u,d\right)\middle|u\right]\right\}. \qquad (1)\end{aligned}$$

We examine each of the three covariances in turn from left to right. Using Lemmas 2 and 3, the first term of (1) is:

$$\mathrm{Cov}\left(\mathrm{E}\left(\hat{\Delta}\middle|u\right),\mathrm{E}\left(\hat{\Psi}\middle|u\right)\right)=\mathrm{Cov}\left(\bar{\Delta},\Delta_2 p_2+(u_2-u_1)\right)=0$$

Now taking the middle term of (1),

$$\mathrm{E}\left\{\mathrm{Cov}\left[\mathrm{E}\left(\hat{\Delta}\middle|u,d\right),\mathrm{E}\left(\hat{\Psi}\middle|u,d\right)\middle|u\right]\right\}$$

$$=\mathrm{E}\left\{\mathrm{Cov}\left[\frac{\Delta_1}{2\hat{q}_2}(Z-\hat{p}''_1)+\bar{\Delta},\right.\right.$$

$$\Delta_2\hat{p}_1''+(u_2-u_1)+\Delta_1(\hat{p}_1''-\hat{p}_1')|u]\}$$
$$=\mathrm{Cov}\left[\frac{\Delta_1}{2\hat{q}_2}(Z-\hat{p}_1''),\Delta_2\hat{p}_2+\Delta_1\hat{p}_1''\right]$$
$$=\frac{\Delta_1}{2}\left[\Delta_2\mathrm{Cov}\left(\frac{Z}{\hat{q}_2},\hat{p}_2\right)+\Delta_1\mathrm{Cov}\left(\frac{Z}{\hat{q}_2},\hat{p}_1''\right)\right]+$$
$$-\Delta_2\mathrm{Cov}\left(\frac{\hat{p}_1''}{\hat{q}_2},\hat{p}_2\right)-\Delta_1\mathrm{Cov}\left(\frac{\hat{p}_1''}{\hat{q}_2},\hat{p}_1''\right)\Bigg]. \qquad (2)$$

Taking each of these terms in turn,

$$\mathrm{Cov}\left(\frac{Z}{\hat{q}_2},\hat{p}_2\right)=\mathrm{E}\left[\mathrm{Cov}\left(\frac{Z}{\hat{q}_2},\hat{p}_2\right)\middle|\{d_2\}\right]+\mathrm{Cov}\left(\frac{p_1}{\hat{q}_2},\hat{p}_2\right)$$
$$=0+p_1\left[\mathrm{E}\left(\frac{\hat{p}_2}{\hat{q}_2}\right)-p_2\mathrm{E}\left(\frac{1}{\hat{q}_2}\right)\right]$$
$$=0+\mathrm{O}(1/n);$$

$$\mathrm{Cov}\left(\frac{Z}{\hat{q}_2},\hat{p}_1''\right)=\mathrm{E}\left[\mathrm{Cov}\left(\frac{Z}{\hat{q}_2},\hat{p}_1''\right)\middle|\{d_2\}\right]+\mathrm{Cov}\left(\frac{p_1}{\hat{q}_2},p_1\right)$$
$$=\mathrm{E}\left[\mathrm{Cov}\left(\frac{\sum_i d_{2i}d_{1i}''}{\hat{q}_2\sum_i d_{2i}},\frac{\sum_i d_{1i}''}{n_2}\right)\middle|\{d_2\}\right]+0$$
$$=\mathrm{E}\left[\frac{\sum_i d_{2i}\mathrm{Var}(d_{1i}'')}{n_2\hat{q}_2\sum_i d_{2i}''}\right]$$
$$=\mathrm{E}\left[\frac{p_1q_1}{n_2\hat{q}_2}\right]=\frac{p_1q_1}{n_2q_2}+\mathrm{O}(1/n);$$

$$\mathrm{Cov}\left(\frac{\hat{p}_1''}{\hat{q}_2},\hat{p}_2\right)=\mathrm{E}\left[\mathrm{Cov}\left(\frac{\hat{p}_1''}{\hat{q}_2},\hat{p}_2\right)\middle|\{d_2\}\right]+\mathrm{Cov}\left(\frac{p_1}{\hat{q}_2},\hat{p}_2\right)$$
$$=0+\mathrm{O}(1/n);$$

and

$$\mathrm{Cov}\left(\frac{\hat{p}_1''}{\hat{q}_2},\hat{p}_1''\right)=\frac{p_1q_1}{n_2q_2}+\mathrm{O}(1/n).$$

So $\mathrm{E}\left\{\mathrm{Cov}\left[\mathrm{E}\left(\hat{\Delta}|u,d\right),\mathrm{E}\left(\hat{\Psi}|u,d\right)\ |u\right]\right\}=0+\mathrm{O}(1/n)$.

Now taking the third term of (1),

$$\mathrm{E}\left\{\mathrm{E}\left[\mathrm{Cov}\left(\hat{\Delta},\hat{\Psi}|u,d\right)\ |u\right]\right\}=$$
$$\mathrm{E}\left[\frac{1}{2}\mathrm{Cov}\left(\frac{\sum_i d_{1i}'e_{1i}}{\sum_i d_{1i}'}+\frac{\sum_i d_{2i}e_{2i}}{\sum_i d_{2i}}-\frac{\sum_i(1-d_{1i}')e_{1i}}{\sum_i(1-d_{1i}')}+\right.\right.$$
$$\left.\left.-\frac{\sum_i(1-d_{2i})e_{2i}}{\sum_i(1-d_{2i})},\sum_i\frac{e_{2i}}{n_2}-\sum_i\frac{e_{1i}}{n_1}\middle|d\right)\right]$$
$$=\frac{1}{2}\mathrm{E}\left[-\mathrm{Cov}\left(\frac{\sum_i d_{1i}'e_{1i}}{\sum_i d_{1i}'},\sum_i\frac{e_{1i}}{n_1}\middle|d\right)+\right.$$
$$+\mathrm{Cov}\left(\frac{\sum_i(1-d_{1i}')e_{1i}}{\sum_i(1-d_{1i}')},\sum_i\frac{e_{1i}}{n_1}\middle|d\right)+$$
$$+\mathrm{Cov}\left(\frac{\sum_i(d_{2i})e_{2i}}{\sum_i d_{2i}},\sum_i\frac{e_{2i}}{n_2}\middle|d\right)+$$
$$\left.-\mathrm{Cov}\left(\frac{\sum_i(1-d_{2i})e_{2i}}{\sum_i(1-d_{2i})},\sum_i\frac{e_{2i}}{n_2}\middle|d\right)\right]$$
$$=\frac{1}{2}\mathrm{E}\left[\left(-\sigma_1^2+\sigma_1^2+\sigma_2^2-\sigma_2^2\right)\right]=0.$$

Having demonstrated that the first and third component of (1) are exactly 0, and the second term is approximately zero, this completes the proof.