

## Simulation to assess varying coefficients and guideline

This runs a series of small simulations where we examine varying beta coefficients over time. In these simulations, we generate data and estimate the guidelines based on that data, and then compare these estimated recommendations to the oracle truth (using our general theorem) to see how our guideline works when it is technically misspecified.

In particular, this document produces Figure 4 and Table 2 in Appendix B.

### Run simulation across range of sigma\_pre

Our initial simulation has varying coefficients for both  $X$  and  $Z$ , along with  $\theta$ . The misspecification gives a reduced recommendation to match as shown on the figure to the right.

```
# Number of simulation replicates per scenario
K = 25 #1000

cor_Xtheta = c( 0.3, 0.6 )

sigma_pre_tests = seq( 0.3, 1.5, by=0.10 )

names(sigma_pre_tests) = sigma_pre_tests
sim_res_main <- map_df( sigma_pre_tests,
  ~ run_scenario( sigma_pre = .,
    beta_theta_0 = c( 0.5, 1.0 ),
    beta_theta_1 = 1.5,
    beta_x_0 = c( 0.6, 1.1 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.3, 0.7 ),
    beta_z_1 = 1.0,
    cor_XZ = 0.5,
    cor_Xtheta = cor_Xtheta,
    K = K ),
    .id = "sigma_pre" ) %>%
  mutate( sigma_pre = as.numeric(sigma_pre) )

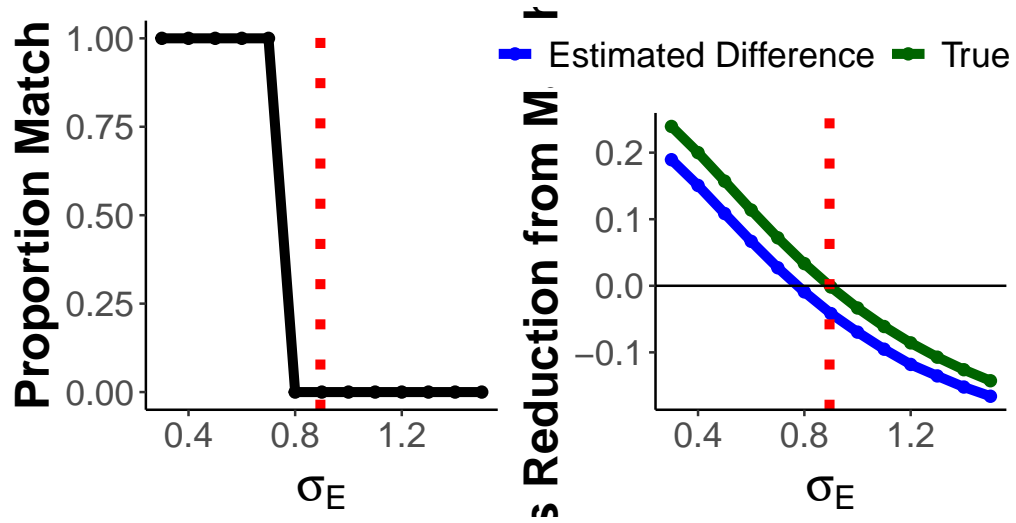
saveRDS(sim_res_main, file = "results/plot_df.rds")
```

Our table shows, for different residual variation, the proportion of the trials that say “match!”, the average estimated reduction in bias, the standard deviation of the estimates across simulation (which is the true SE), and whether the oracle says to match and how much bias would be reduced. The last column is the  $R^2$  for a regression of outcome onto the two observed covariates to get a sense of how much variation is explained by what we can match on.

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.3	1	0.189	0.007	TRUE	0.239	0.833
0.4	1	0.151	0.007	TRUE	0.200	0.813
0.5	1	0.109	0.004	TRUE	0.157	0.790

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.6	1	0.067	0.005	TRUE	0.114	0.761
0.7	1	0.027	0.005	TRUE	0.072	0.735
0.8	0	-0.009	0.004	TRUE	0.033	0.703
0.9	0	-0.042	0.005	FALSE	-0.002	0.673
1.0	0	-0.070	0.006	FALSE	-0.033	0.644
1.1	0	-0.095	0.008	FALSE	-0.061	0.614
1.2	0	-0.118	0.009	FALSE	-0.086	0.587
1.3	0	-0.135	0.010	FALSE	-0.107	0.557
1.4	0	-0.152	0.011	FALSE	-0.126	0.532
1.5	0	-0.166	0.014	FALSE	-0.143	0.506

The plot is as discussed in the appendix:

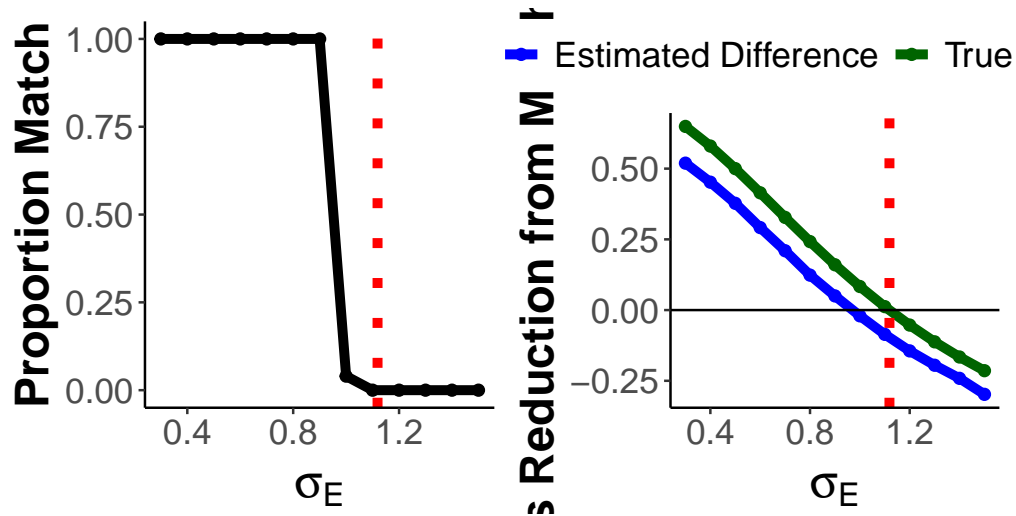


## Independent covariates

This is our initial simulation, above, except all three covariates are now independent from each other. The independence does not help us?

```
sim_res_indep <- map_df( sigma_pre_tests,
  ~ run_scenario( sigma_pre = .,
    beta_theta_0 = c( 0.5, 1.0 ),
    beta_theta_1 = 1.5,
    beta_x_0 = c( 0.6, 1.1 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.3, 0.7 ),
    beta_z_1 = 1.0,
    cor_XZ = 0.0,
    cor_Xtheta = c(0,0),
    K = K ),
    .id = "sigma_pre" ) %>%
  mutate( sigma_pre = as.numeric(sigma_pre) )
```

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.3	1.00	0.519	0.012	TRUE	0.649	0.567
0.4	1.00	0.452	0.009	TRUE	0.580	0.541
0.5	1.00	0.377	0.009	TRUE	0.500	0.519
0.6	1.00	0.292	0.011	TRUE	0.415	0.488
0.7	1.00	0.210	0.008	TRUE	0.328	0.460
0.8	1.00	0.123	0.010	TRUE	0.242	0.432
0.9	1.00	0.050	0.010	TRUE	0.160	0.406
1.0	0.04	-0.020	0.012	TRUE	0.083	0.375
1.1	0.00	-0.086	0.013	TRUE	0.012	0.351
1.2	0.00	-0.144	0.014	FALSE	-0.053	0.328
1.3	0.00	-0.195	0.015	FALSE	-0.112	0.309
1.4	0.00	-0.241	0.018	FALSE	-0.166	0.288
1.5	0.00	-0.298	0.019	FALSE	-0.214	0.269



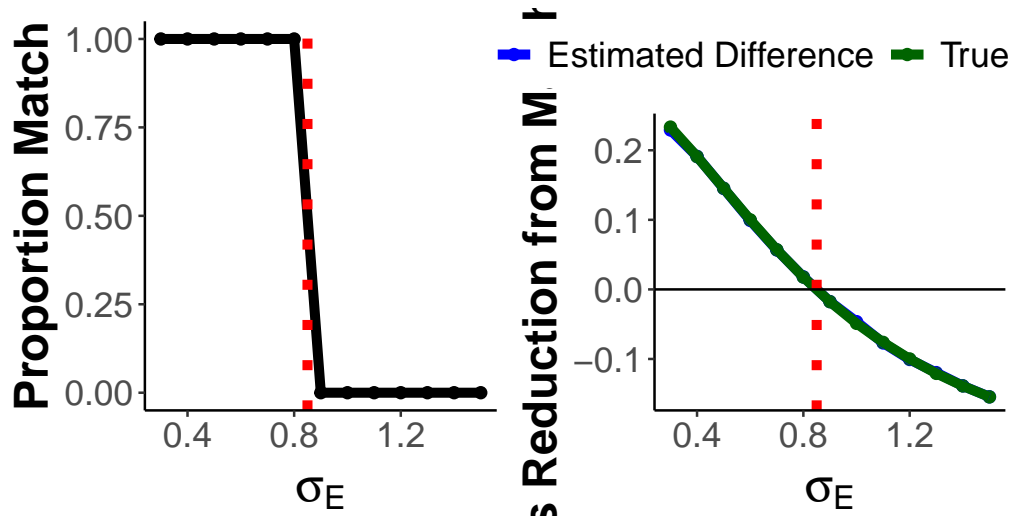
## All is parallel

If we have parallel trends for all three covariates (two observed, one latent) then our guideline works as expected.

Also note that due to large sample size our match recommendation is very precisely estimated because our bias reduction is also very precisely estimated (see the SE column).

```
sim_res_correct <- map_df( sigma_pre_tests,
  ~ run_scenario( sigma_pre = .,
    K = K,
    beta_theta_0 = c( 0.75, 0.75 ),
    beta_theta_1 = 1.5,
    beta_x_0 = c( 0.85, 0.85 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.5, 0.5 ),
    beta_z_1 = 1.0,
    cor_Xtheta = cor_Xtheta,
    cor_XZ = 0.5 ),
  mutate( sigma_pre = as.numeric(sigma_pre) ) )
```

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.3	1	0.229	0.009	TRUE	0.233	0.840
0.4	1	0.191	0.008	TRUE	0.191	0.824
0.5	1	0.145	0.008	TRUE	0.145	0.806
0.6	1	0.099	0.005	TRUE	0.100	0.783
0.7	1	0.057	0.005	TRUE	0.057	0.760
0.8	1	0.018	0.005	TRUE	0.018	0.734
0.9	0	-0.018	0.004	FALSE	-0.018	0.706
1.0	0	-0.046	0.006	FALSE	-0.049	0.678
1.1	0	-0.077	0.008	FALSE	-0.076	0.650
1.2	0	-0.101	0.009	FALSE	-0.100	0.621
1.3	0	-0.120	0.008	FALSE	-0.121	0.595
1.4	0	-0.139	0.011	FALSE	-0.139	0.566
1.5	0	-0.154	0.013	FALSE	-0.155	0.541

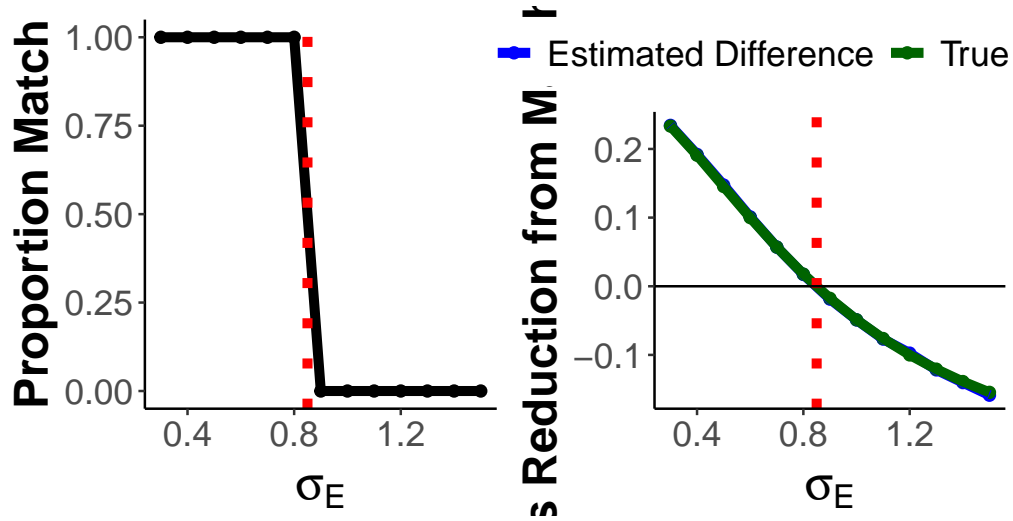


## theta parallel, covariates not

Here we have X and Z correlated, but theta is parallel.

```
sim_res_theta_par <- map_df( sigma_pre_tests,
  ~ run_scenario( sigma_pre = .,
    K = K,
    beta_theta_0 = c( 0.75, 0.75 ),
    beta_theta_1 = 1.5,
    beta_x_0 = c( 0.6, 1.1 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.3, 0.7 ),
    beta_z_1 = 1.0,
    cor_Xtheta = cor_Xtheta,
    cor_XZ = 0.5 ),
    .id = "sigma_pre" ) %>%
mutate( sigma_pre = as.numeric(sigma_pre) )
```

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.3	1	0.234	0.010	TRUE	0.233	0.825
0.4	1	0.192	0.009	TRUE	0.191	0.808
0.5	1	0.147	0.006	TRUE	0.145	0.790
0.6	1	0.101	0.007	TRUE	0.100	0.765
0.7	1	0.057	0.004	TRUE	0.057	0.740
0.8	1	0.017	0.005	TRUE	0.018	0.713
0.9	0	-0.019	0.005	FALSE	-0.018	0.686
1.0	0	-0.049	0.006	FALSE	-0.049	0.657
1.1	0	-0.077	0.008	FALSE	-0.076	0.630
1.2	0	-0.097	0.008	FALSE	-0.100	0.602
1.3	0	-0.122	0.009	FALSE	-0.121	0.574
1.4	0	-0.140	0.010	FALSE	-0.139	0.547
1.5	0	-0.159	0.015	FALSE	-0.155	0.523

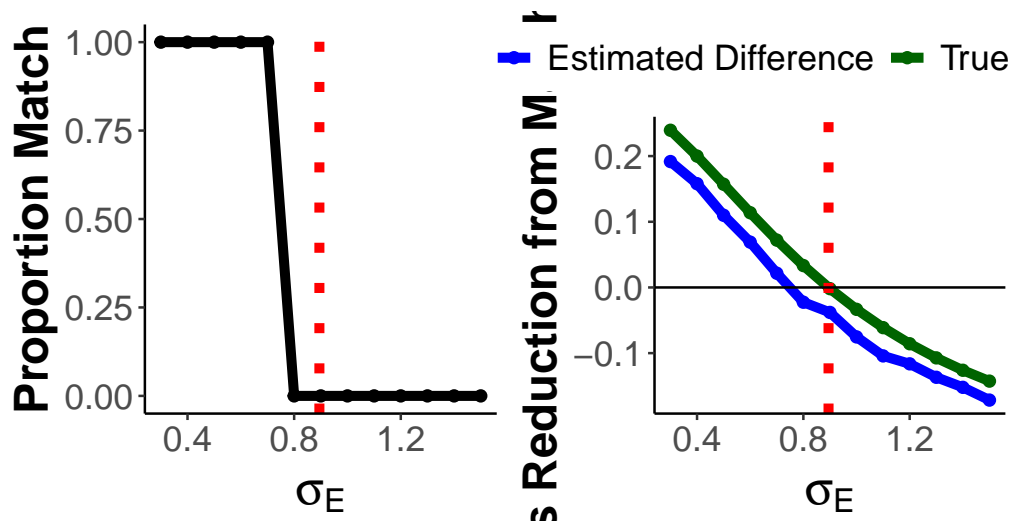


## small sample size

From our original simulation, if we reduce sample size, estimation error should flatten our curve. It seems like estimation error is very small, which is surprising given all the residualization?

```
sim_resSS <- map_df( sigma_pre_tests,
  ~ run_scenario( sigma_pre = .,
    beta_theta_0 = c( 0.5, 1.0 ),
    beta_theta_1 = 1.5,
    beta_x_0 = c( 0.6, 1.1 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.3, 0.7 ),
    beta_z_1 = 1.0,
    cor_Xtheta = cor_Xtheta,
    cor_XZ = 0.5,
    N = 3000 ),
    .id = "sigma_pre" ) %>%
  mutate( sigma_pre = as.numeric(sigma_pre) )
```

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.3	1	0.192	0.024	TRUE	0.239	0.833
0.4	1	0.158	0.016	TRUE	0.200	0.814
0.5	1	0.110	0.010	TRUE	0.157	0.785
0.6	1	0.069	0.013	TRUE	0.114	0.756
0.7	1	0.022	0.006	TRUE	0.072	0.735
0.8	0	-0.022	0.012	TRUE	0.033	0.703
0.9	0	-0.038	0.011	FALSE	-0.002	0.669
1.0	0	-0.075	0.012	FALSE	-0.033	0.644
1.1	0	-0.104	0.022	FALSE	-0.061	0.610
1.2	0	-0.116	0.027	FALSE	-0.086	0.584
1.3	0	-0.137	0.020	FALSE	-0.107	0.555
1.4	0	-0.152	0.023	FALSE	-0.126	0.532
1.5	0	-0.171	0.034	FALSE	-0.143	0.509



The uncertainty in estimation of the reduction does go up quite a bit:

```
sim_resSS$SE_tau_xy / sim_res_main$SE_tau_xy
```

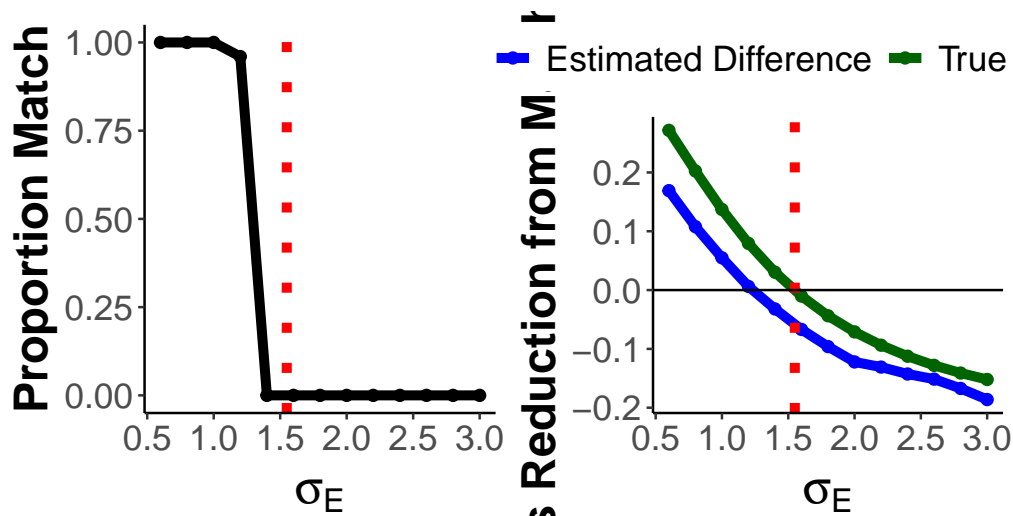
```
## [1] 3.577619 2.268666 2.697294 2.673164 1.352320 3.122661 2.107932 1.894204  
## [9] 2.667425 2.872675 1.966691 2.018256 2.369015
```

## More time periods with less predictive covariates

If we make covariates less predictive, but have more pre-treatment time periods? (We have also increased variation in theta to cause more trouble.)

```
sigmas_larger = sigma_pre_tests * 2
names(sigmas_larger) = sigmas_larger
sim_res_T4 <- map_df( sigmas_larger,
  ~ run_scenario( sigma_pre = .,
    beta_theta_0 = c( 0.0, 0.4, 0.8, 1.2 ),
    beta_theta_1 = 1.6,
    beta_x_0 = c( 0.6, 1.1, 1.1, 0.6 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.7, 0.7, 0.3, 0.3 ),
    beta_z_1 = 1.0,
    cor_Xtheta = cor_Xtheta,
    cor_XZ = 0.5,
    K = K ),
  .id = "sigma_pre" ) %>%
mutate( sigma_pre = as.numeric(sigma_pre) )
```

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.6	1.00	0.169	0.008	TRUE	0.272	0.759
0.8	1.00	0.108	0.005	TRUE	0.202	0.696
1.0	1.00	0.055	0.005	TRUE	0.137	0.632
1.2	0.96	0.006	0.005	TRUE	0.079	0.570
1.4	0.00	-0.032	0.007	TRUE	0.030	0.510
1.6	0.00	-0.067	0.009	FALSE	-0.010	0.459
1.8	0.00	-0.096	0.013	FALSE	-0.044	0.409
2.0	0.00	-0.122	0.013	FALSE	-0.071	0.366
2.2	0.00	-0.131	0.019	FALSE	-0.094	0.332
2.4	0.00	-0.143	0.019	FALSE	-0.112	0.300
2.6	0.00	-0.151	0.018	FALSE	-0.128	0.272
2.8	0.00	-0.167	0.025	FALSE	-0.141	0.247
3.0	0.00	-0.186	0.026	FALSE	-0.152	0.223



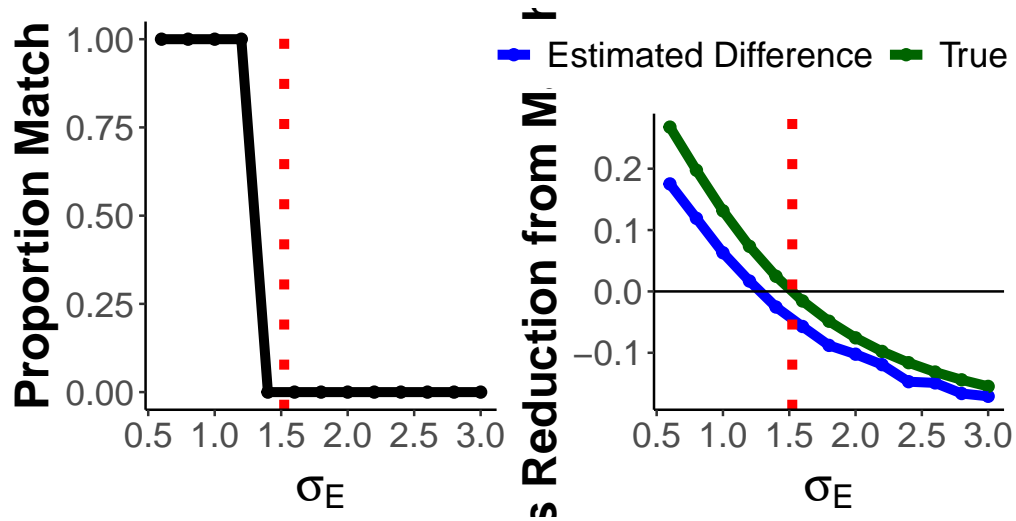


## narrow theta assumption only

If all we get is parallel theta in the final two periods, but theta is not parallel before that?

```
sim_res_T4_narrow <- map_df( sigmas_larger,
  ~ run_scenario( sigma_pre = .,
    beta_theta_0 = c( 0.0, 0.4, 1.0, 1.0 ),
    beta_theta_1 = 1.6,
    beta_x_0 = c( 0.6, 1.1, 1.1, 0.6 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.7, 0.7, 0.3, 0.3 ),
    beta_z_1 = 1.0,
    cor_Xtheta = cor_Xtheta,
    cor_XZ = 0.5,
    K = K ),
  .id = "sigma_pre" ) %>%
mutate( sigma_pre = as.numeric(sigma_pre) )
```

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.6	1	0.175	0.009	TRUE	0.268	0.759
0.8	1	0.119	0.006	TRUE	0.197	0.696
1.0	1	0.063	0.003	TRUE	0.131	0.632
1.2	1	0.017	0.005	TRUE	0.073	0.568
1.4	0	-0.025	0.005	TRUE	0.025	0.510
1.6	0	-0.057	0.008	FALSE	-0.016	0.456
1.8	0	-0.088	0.012	FALSE	-0.049	0.411
2.0	0	-0.102	0.012	FALSE	-0.076	0.368
2.2	0	-0.119	0.011	FALSE	-0.098	0.333
2.4	0	-0.147	0.019	FALSE	-0.116	0.298
2.6	0	-0.149	0.024	FALSE	-0.131	0.271
2.8	0	-0.166	0.014	FALSE	-0.144	0.245
3.0	0	-0.171	0.017	FALSE	-0.155	0.222

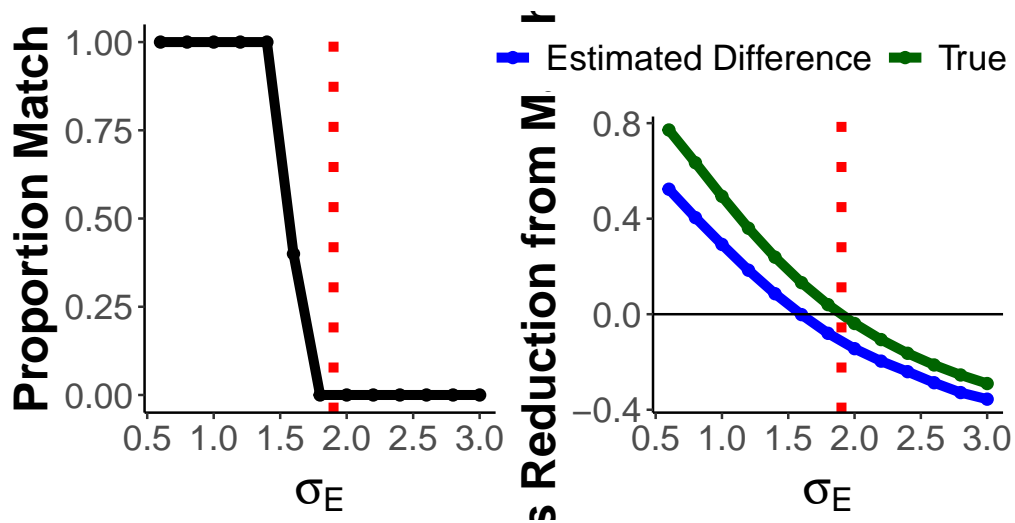


## All covariates independent, four time points

Here we have our theta stable in the final two time periods, and no correlation between any of our three covariates.

```
sim_res_T4_indep <- map_df( sigmas_larger,
  ~ run_scenario( sigma_pre = .,
    beta_theta_0 = c( 0.0, 0.4, 1.0, 1.0 ),
    beta_theta_1 = 1.6,
    beta_x_0 = c( 0.6, 1.1, 1.1, 0.6 ),
    beta_x_1 = 1.3,
    beta_z_0 = c( 0.7, 0.7, 0.3, 0.3 ),
    beta_z_1 = 1.0,
    cor_XZ = 0,
    cor_Xtheta = c( 0, 0 ),
    K = K ),
  .id = "sigma_pre" ) %>%
mutate( sigma_pre = as.numeric(sigma_pre) )
```

sigma_pre	per_match	a_tau_xy	SE_tau_xy	match_XY	reduce_XY	R2
0.6	1.0	0.523	0.010	TRUE	0.771	0.499
0.8	1.0	0.404	0.013	TRUE	0.634	0.438
1.0	1.0	0.292	0.010	TRUE	0.494	0.380
1.2	1.0	0.184	0.012	TRUE	0.360	0.330
1.4	1.0	0.086	0.009	TRUE	0.239	0.285
1.6	0.4	-0.002	0.012	TRUE	0.132	0.247
1.8	0.0	-0.080	0.012	TRUE	0.040	0.216
2.0	0.0	-0.144	0.019	FALSE	-0.039	0.189
2.2	0.0	-0.196	0.022	FALSE	-0.106	0.166
2.4	0.0	-0.241	0.025	FALSE	-0.164	0.147
2.6	0.0	-0.287	0.026	FALSE	-0.213	0.131
2.8	0.0	-0.328	0.028	FALSE	-0.254	0.118
3.0	0.0	-0.356	0.032	FALSE	-0.290	0.106



## All biases across scenarios

These plots aggregate the above to compare trends and sizes of biases, etc.

