Appendix B: Power Analysis

B.1 Simulated Power Analysis Method

To estimate the minimum detectable effect size on the fertility of abortion bans at various levels of power, we follow the methods presented by Black et al. (2022).

Like Black et al. (2022) we assign a pseudo period like the Dobbs period in the main analysis. For the main results of our paper, we rely on the 4 years leading into Dobbs in the pre-period and 1-year post, which is 2019-2022 and 2023 respectively for the Dobbs period. For the pseudo period, we rely on 2015-2019. We set the period of analysis to 2015-2019 to exclude any actual effects in differential fertility that might have occurred due to the COVID-19 pandemic across states (Bailey et al., 2022; Kearney and Levine, 2023; Dench et al., 2023). We randomly assign treatment to 12 states to match the number of states with bans going into effect shortly after *Dobbs* but excluding Texas for the reasons discussed in the main body of the paper. We then impose varying effects starting from the null and increasing out to 7 percent positive and 7 percent negative effects of the mean fertility rate in each population in whole percentage point increments on the last year of the pseudo-treatment period, 2019. We estimate Synthetic difference-in-differences and two-way fixed effects models where pseudo-treatment turns on in 2019. For Synthetic difference-in-differences models, we use cluster bootstrap standard errors with 1,000 bootstrap samples. For two-way fixed models we cluster standard errors at the state level. Then, we repeat this randomization and analysis 200 times and report the percent of samples at each effect size where we have t-statistics either greater than 1.96 or less than -1.96, representing the power of the test at that effect size corresponding to a rejection rate of 0.05 on a two-sided hypothesis.

Our method differs from Black et al. (2022) in the following ways. First, in some analysis Black et al. (2022) adjust the weighting of their analysis by applying inverse propensity weighting based on observable characteristics (IPW) so that the weighted randomized pseudo-treated look more like the set of groups that are actually treated and the randomized pseudo-control groups look more like the set of groups that are actually control groups.

In our case, we do not think it is reasonable to apply this adjustment since Dobbs states will vary on many unobservable as well as observable characteristics that would make such reweighting implausible. In addition, Synthetic difference-in-differences (SDID), adjusts for the most obvious and important difference between treated and control which is differential trends, without the need to arbitrarily select reweighting control variables to meet this condition. Second, we consider a two-tailed instead of a one-tailed test. While we believe treatment effects should be positive, given the substantial literature in support of this, we do not want to impose that given the uncertainty around *Dobbs'* effects on mitigating behaviors. Second, because this is a state-level analysis, there is no need to randomly increase births such as Black et al. (2022) removed deaths based on their probability of occurring in each county. Instead, we simply, increase or decrease the fertility rate by the selected percent of the state-year fertility rates. Third, we do not remove states where there could be pre-treatment contamination from the analysis. Given ban states are frequent regulators of abortion, we would have very few treatment states from which to draw inference in randomization. Instead, we rely on the parallel trends assumption inherent in difference-in-differences which holds on average under randomization.

To assess sensitivity to the selection of pre-period, in another set of power analyses, we lengthen our pre-period to go back to 2005. This is to show the extent to which power may be affected in two-way fixed effects or synthetic-difference-in-differences by arbitrary selection of the pre-period. The caveat is that synthetic differences in differences might still select a weighted set of pre-periods. In that case there is no subjective judgment in selection but rather it is based on the algorithm defined in Arkhangelsky et al. (2021). Randomization in this case will still impose parallel trends on average.

Our power analysis illustrates the effect sizes we could detect if treatment were randomized across states. It should be noted that randomization imposes the assumption, on average, that the treated group is trending similarly to the control group. It therefore guarantees the underlying assumption in difference-in-differences analysis while this might not play out in the real world if there are differential trends between treatment and control groups. These power calculations should therefore be considered in the context of where

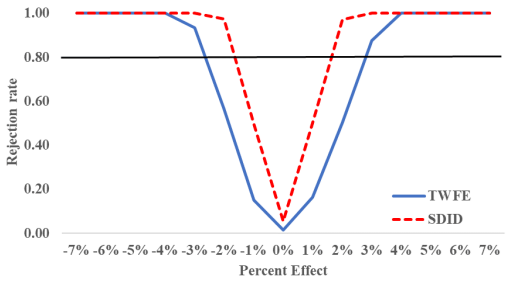
this underlying assumption holds.

To address this concern, we conduct a secondary analysis assessing effects assigned to the *Dobbs* ban states over time. We reassign the period of analysis to every 5-year period from 2005 to 2019 (e.g. 2005-2009, 2006-2010), generating eleven pseudo-periods, imposing the effects in the last year of treatment. The rest of the power analysis follows similarly.

B.2 Power analysis results

Our primary power analysis in table B.1 and figure B.1 to B.3 shows the rejection rate imposing each randomized treatment effect of between -7 to 7 percent limited to the period from 2015-2019. We report ranges for MDE at conventional levels of 0.8 power levels and only on the positive side. In the case of randomization, however, positive and negative rejection MDE are quite symmetric. As expected, the rejection rate when there is zero imposed effect is at or around 0.05 for both methods due to randomization. The main difference between TWFE and SDID is that SDID achieves the conventional rate of rejection of 80% or more much more quickly, both overall and in each of our subpopulations of interest. For the overall population synthetic difference-in-differences reaches the conventional power level between 1-2 percent imposed effects, whereas TWFE reaches this level between 2-3 percent imposed effects. To be more specific, SDID, the more powerful method, crosses the 80% threshold between 1.4 to 1.6% effects. For the age group, 15-19, TWFE reaches the conventional power level after 7 percent imposed effects, whereas SDID reaches the conventional power level between 5 to 6 percent. For 20-24, the conventional power level is reached between 3 to 4 percent but 2 to 3 percent for SDID. For age 25-29 TWFE hits the conventional power level at between 4 and 5 percent, whereas SDID hits that level between 2 to 3 percent. For non-Hispanic white TWFE hits the conventional power level between 2 to 3 percent whereas for TWFE the MDE is between 1 to 2 percent. For non-hispanic black we hit the conventional level closer to six percent, while for SDID we hit the conventional level closer to five percent. Finally, for Hispanic women, we cross the conventional power level between 4 to 5 percent for TWFE and 3 to 4 percent for SDID.

In table B.2 and figure B.4 to B.6 we report how extending the pre-period of analysis

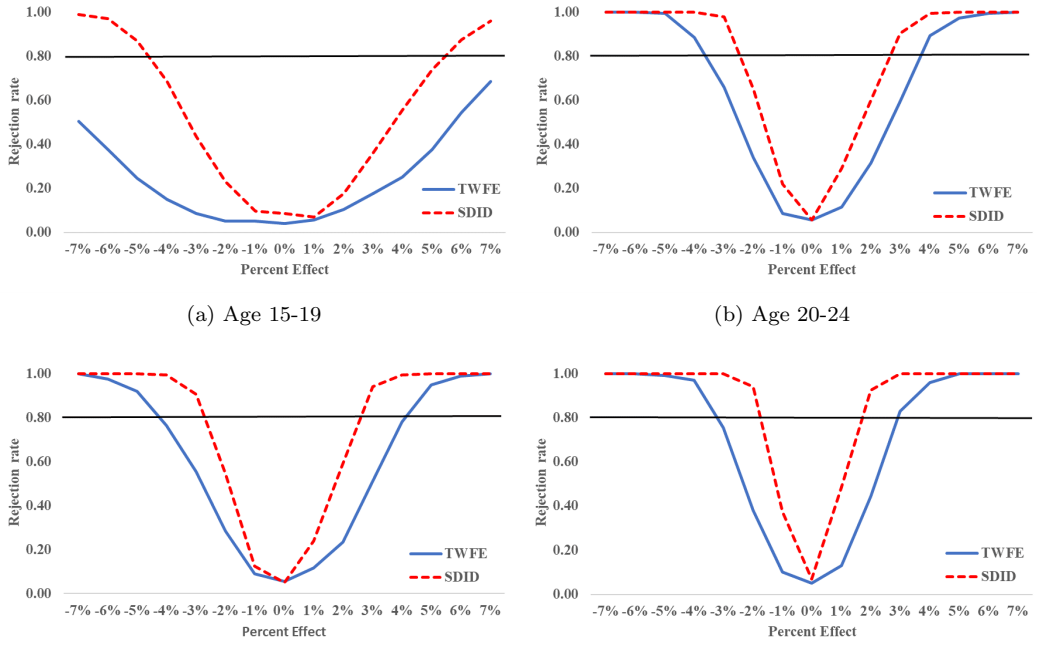


(a) All women

Figure B.1: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on the period 2015-2019 in a random set of 12 states that mimic the bans in the twelve *Dobbs* ban states excluding Texas for the overall population. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.

effects MDE's for each group of interest. The MDE are all higher in the case of OLS, for some groups substantially so, but practically unchanged for SDID. This is likely because of SDID's automatic selection of time-weights to reduce the difference in the average post-period and pre-period for the control group. In this way, for power, SDID is rather insensitive to selection of the pre-period.

The results in **??** and B.7 to B.9 show the result of imposing an effect on the Dobbs states at the end of each set of five-year periods from 2005-2019. Keep in mind that this is more akin to a placebo in time analysis than a simulated power analysis in the spirit of Black et al. (2022). If there are any actual differential effects between Dobbs states and non-Dobbs states during the last year of these periods or non-parallel trends in the pre-period, than it will contaminate the analysis and create skewed overrejection in either the positive or negative direction. Also, keep in mind that there are 11 potential time periods, so there is potentially substantial sampling variance, and thus the results of this power analysis will inherently contain more noise than the randomization analysis. What we see is that for all women, age 15-44, for both TWFE and SDID, you are still very unlikely to reject the null hypothesis around the null. Like with randomization power analysis in B.1 you get to



(a) Age 15-19

(b) Age 20-24

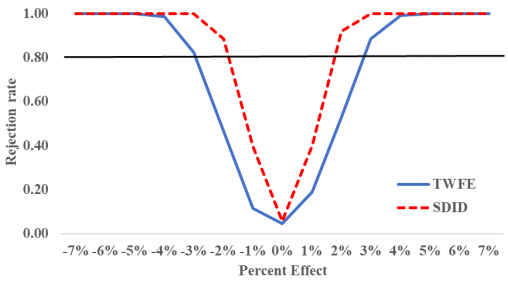
(c) Age 25-29

(d) Age 30

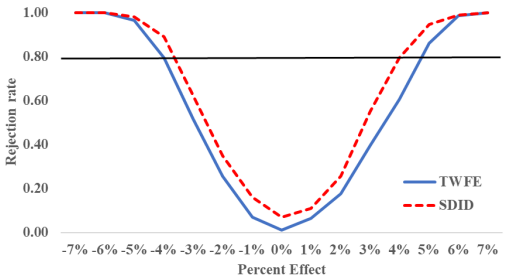
Figure B.2: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on the period 2015-2019 in a random set of 12 states that mimic the bans in the twelve *Dobbs* ban states excluding Texas by age group. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.

conventionally powered levels around 1 to 2 percent for SDID and 2 to 3 for TWFE. This is because the trends in fertility over the entire period 2005-2019 for women age 15-44 in the Dobbs states and non-Dobbs states were parallel as can be seen in Appendix **??** in TWFE and SDID event studies.

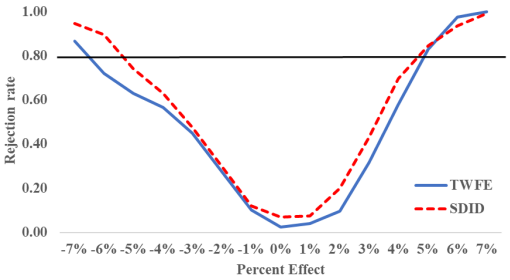
In the case of women-age 15-19, you reject the null hypothesis in TWFE with zero imposed effects 55% of the time. This is not unexpected given the trends observed in TWFE event studies in Appendix 6 in the pre-period. We also fail to ever reject the null hypothesis for age 15-19 women using TWFE on the positive end. By contrast, in SDID you reject the null with zero imposed effect only two out of 11 times and can detect effects on



(a) White Non-Hispanic



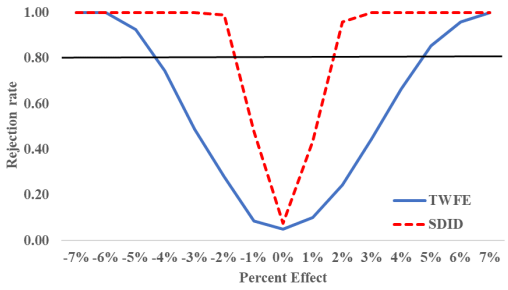
(b) Black Non-Hispanic



(c) Hispanic

Figure B.3: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on the period 2015-2019 in a random set of 12 states that mimic the bans in the twelve *Dobbs* ban states excluding Texas by race group. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.

both sides of the distribution. With positive imposed effects, you reject the null hypothesis at greater than an 80% rate between 5 to 6 percent which is similar to the rejection rates in the randomization power analysis. In all age groups, rejections are more symmetric in SDID than in TWFE and in almost all cases where there is not severe skew in one direction with TWFE, conventional power rates are crossed earlier for SDID than for TWFE. One other thing to note is that, as is the case for age group 25-29 as seen in related event studies, when no weighted set of control states have similar trends to the treatment group, rejection at the null is quite common and may hinder our ability to interpret causality in this group. It should be noted that rejection at the null is also very high for TWFE.



(a) All women

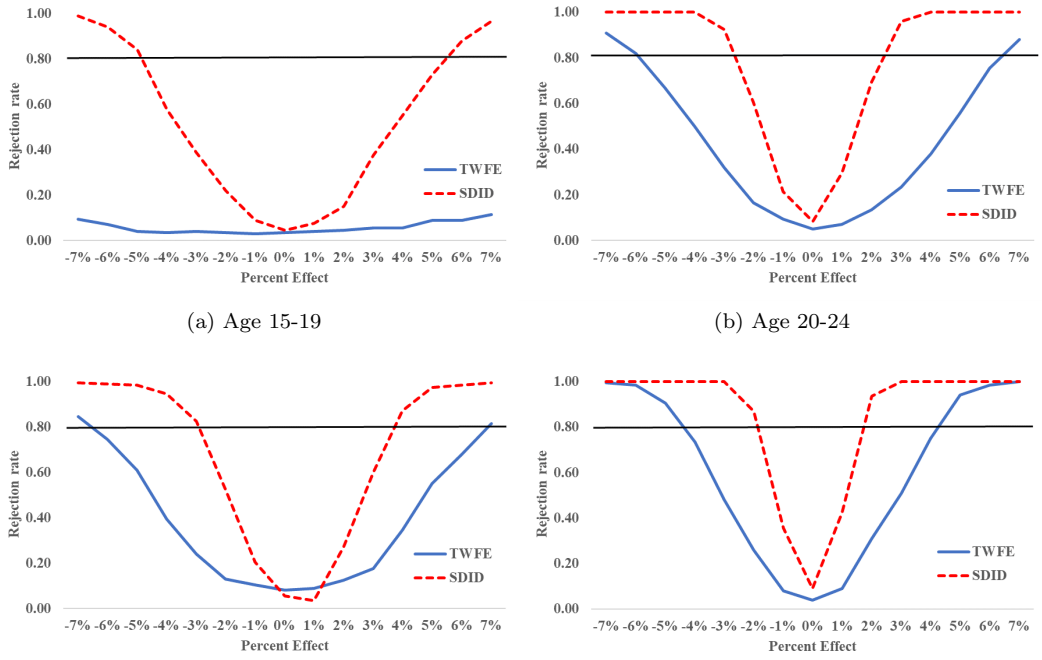
Figure B.4: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on the period 2005-2019 in a random set of 12 states that mimic the bans in the twelve *Dobbs* ban states excluding Texas for the overall population. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.

Table B.1: Two-way fixed effects versus Synthetic difference-indifferences simulated power analysis rejection rates on randomly imposing Dobbs, 2015-2019

Percent Effect	Overall	Age 15-19	Age 20-24	Age 25-29	Age 30+	White	Black	Hispanic
Two-way Fixed Effects								
-7	1.00	0.51	1.00	1.00	1.00	1.00	0.87	1.00
-6	1.00	0.38	1.00	0.98	1.00	1.00	0.72	1.00
-5	1.00	0.25	1.00	0.92	1.00	0.99	0.63	0.97
-4	1.00	0.15	0.89	0.77	0.99	0.95	0.45	0.80
-3	0.94	0.09	0.66	0.56	0.83	0.76	0.45	0.52
-2	0.57	0.05	0.34	0.29	0.47	0.38	0.28	0.26
-1	0.15	0.05	0.09	0.09	0.12	0.10	0.10	0.07
0	0.02	0.04	0.06	0.06	0.05	0.05	0.03	0.01
1	0.17	0.06	0.12	0.12	0.19	0.13	0.04	0.07
2	0.51	0.11	0.32	0.24	0.53	0.45	0.1	0.18
3	0.88	0.18	0.60	0.51	0.89	0.83	0.32	0.40
4	1.00	0.25	0.90	0.78	0.99	0.96	0.58	0.61
5	1.00	0.38	0.98	0.95	1.00	1.00	0.83	0.86
6	1.00	0.55	1.00	0.99	1.00	1.00	0.98	0.99
7	1.00	0.69	1.00	1.00	1.00	1.00	1.00	1.00
Synthetic Difference-in-differences								
-7	1.00	0.99	1.00	1.00	1.00	1.00	0.94	1.00
-6	1.00	0.97	1.00	1.00	1.00	1.00	0.89	1.00
-5	1.00	0.87	1.00	1.00	1.00	1.00	0.74	0.98
-4	1.00	0.69	1.00	1.00	1.00	1.00	0.63	0.89
-3	1.00	0.44	0.98	0.91	1.00	1.00	0.48	0.62
-2	0.98	0.23	0.65	0.55	0.89	0.95	0.30	0.35
-1	0.50	0.10	0.22	0.13	0.40	0.38	0.12	0.16
0	0.06	0.09	0.06	0.05	0.06	0.07	0.07	0.07
1	0.50	0.07	0.29	0.24	0.40	0.49	0.08	0.11
2	0.97	0.18	0.60	0.59	0.92	0.93	0.20	0.26
3	1.00	0.36	0.91	0.94	1.00	1.00	0.43	0.55
4	1.00	0.56	1.00	1.00	1.00	1.00	0.69	0.80
5	1.00	0.74	1.00	1.00	1.00	1.00	0.84	0.95
6	1.00	0.88	1.00	1.00	1.00	1.00	0.94	0.99
7	1.00	0.96	1.00	1.00	1.00	1.00	0.99	1.00

Table B.2: Two-way fixed effects versus Synthetic difference-indifferences simulated power analysis rejection rates on randomly imposing Dobbs, 2005-2019

Percent Effect	Overall	Age 15-19	Age 20-24	Age 25-29	Age 30+	Hispanic
Two-way Fixed Effects						
-7	1.00	0.10	0.85	0.91	1.00	0.42
-6	1.00	0.07	0.75	0.82	0.99	0.33
-5	0.93	0.04	0.61	0.67	0.94	0.27
-4	0.75	0.04	0.40	0.50	0.76	0.19
-3	0.49	0.04	0.24	0.32	0.47	0.13
-2	0.28	0.04	0.13	0.17	0.22	0.07
-1	0.09	0.03	0.11	0.10	0.09	0.07
0	0.05	0.04	0.08	0.05	0.06	0.05
1	0.10	0.04	0.09	0.07	0.14	0.06
2	0.25	0.05	0.13	0.14	0.31	0.07
3	0.45	0.06	0.18	0.24	0.58	0.11
4	0.67	0.06	0.35	0.38	0.82	0.13
5	0.86	0.09	0.55	0.56	0.94	0.17
6	0.96	0.09	0.68	0.76	0.99	0.24
7	1.00	0.12	0.82	0.88	1.00	0.31
Synthetic Difference-in-differences						
-7	1.00	0.99	1.00	1.00	1.00	1.00
-6	1.00	0.94	0.99	1.00	1.00	0.99
-5	1.00	0.84	0.99	1.00	1.00	0.97
-4	1.00	0.58	0.95	1.00	1.00	0.87
-3	1.00	0.39	0.83	0.93	1.00	0.64
-2	0.99	0.22	0.53	0.61	0.96	0.30
-1	0.48	0.09	0.21	0.22	0.41	0.11
0	0.08	0.05	0.06	0.09	0.03	0.06
1	0.44	0.08	0.04	0.30	0.34	0.13
2	0.96	0.15	0.27	0.70	0.91	0.39
3	1.00	0.38	0.60	0.96	1.00	0.67
4	1.00	0.55	0.87	1.00	1.00	0.85
5	1.00	0.73	0.98	1.00	1.00	0.97
6	1.00	0.88	0.99	1.00	1.00	1.00
7	1.00	0.97	1.00	1.00	1.00	1.00



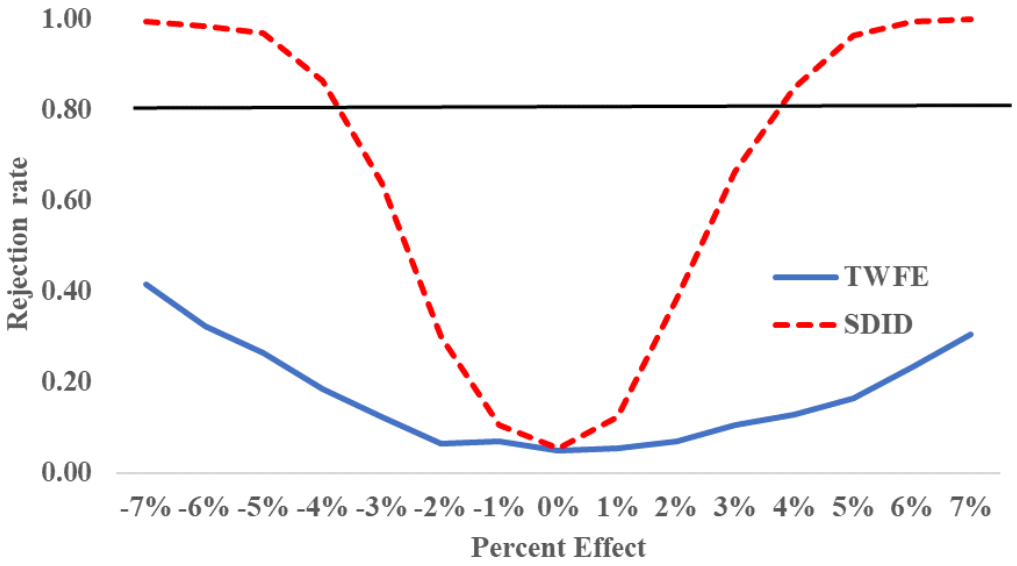
(a) Age 15-19

(b) Age 20-24

(c) Age 25-29

(d) Age 30

Figure B.5: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on the period 2005-2019 in a random set of 12 states that mimic the bans in the twelve *Dobbs* ban states excluding Texas by age group. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.

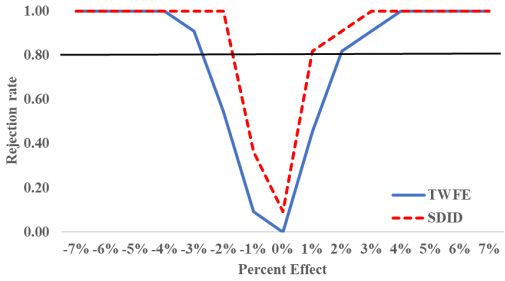


(a) Hispanic

Figure B.6: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on the period 2005-2019 in a random set of 12 states that mimic the bans in the twelve *Dobbs* ban states excluding Texas for Hispanic women. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.

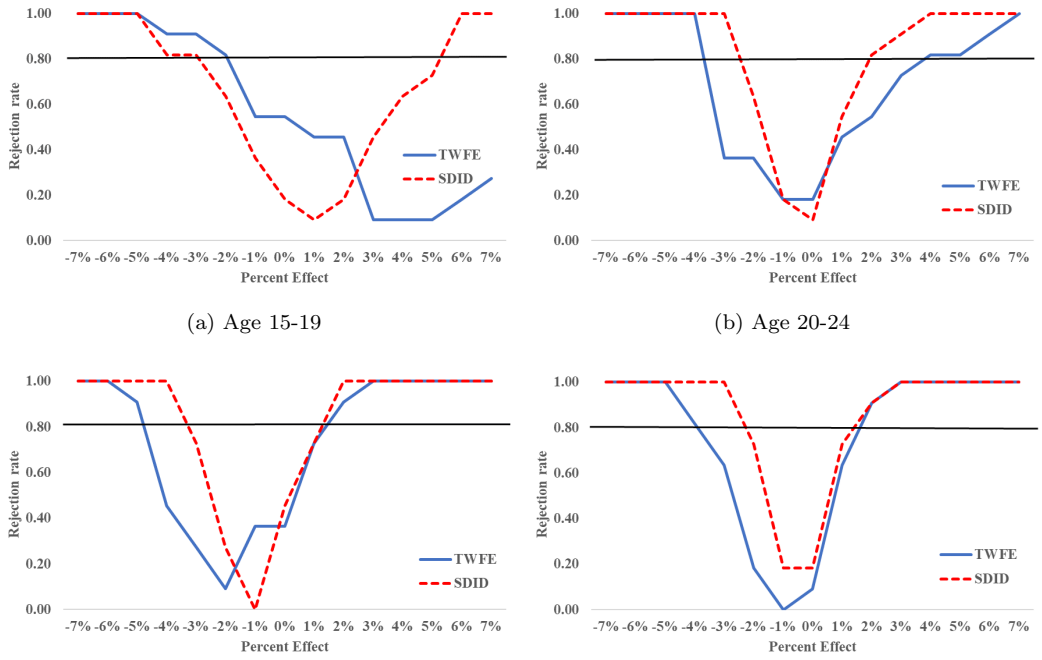
Table B.3: Two-way fixed effects versus Synthetic difference-in-differences simulated power analysis rejection rates on randomly imposing Dobbs in different time period, 2005-2019

Percent Effect	Overall	Age 15-19	Age 20-24	Age 25-29	Age 30+	Hispanic
Two-way Fixed Effects						
-7	1.00	1.00	1.00	1.00	1.00	0.73
-6	1.00	1.00	1.00	1.00	1.00	0.73
-5	1.00	1.00	1.00	0.91	1.00	0.64
-4	1.00	0.91	1.00	0.45	0.82	0.64
-3	0.91	0.91	0.36	0.27	0.64	0.18
-2	0.55	0.82	0.36	0.09	0.18	0.18
-1	0.09	0.55	0.18	0.36	0.00	0.18
0	0.00	0.55	0.18	0.36	0.09	0.18
1	0.45	0.45	0.45	0.73	0.64	0.18
2	0.82	0.45	0.55	0.91	0.91	0.36
3	0.91	0.09	0.73	1.00	1.00	0.55
4	1.00	0.09	0.82	1.00	1.00	0.55
5	1.00	0.09	0.82	1.00	1.00	0.55
6	1.00	0.18	0.91	1.00	1.00	0.55
7	1.00	0.27	1.00	1.00	1.00	0.64
Synthetic Difference-in-differences						
-7	1.00	1.00	1.00	1.00	1.00	0.91
-6	1.00	1.00	1.00	1.00	1.00	0.82
-5	1.00	1.00	1.00	1.00	1.00	0.73
-4	1.00	0.82	1.00	1.00	1.00	0.55
-3	1.00	0.82	1.00	0.73	1.00	0.45
-2	1.00	0.64	0.64	0.27	0.73	0.36
-1	0.36	0.36	0.18	0.00	0.18	0.18
0	0.09	0.18	0.09	0.45	0.18	0.18
1	0.82	0.09	0.55	0.73	0.73	0.18
2	0.91	0.18	0.82	1.00	0.91	0.27
3	1.00	0.45	0.91	1.00	1.00	0.64
4	1.00	0.64	1.00	1.00	1.00	0.73
5	1.00	0.73	1.00	1.00	1.00	0.82
6	1.00	1.00	1.00	1.00	1.00	0.91
7	1.00	1.00	1.00	1.00	1.00	0.91



(a) All women

Figure B.7: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on 12 states excluding Texas at the end of every 5-year period from 2005-2019 for the overall population. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.



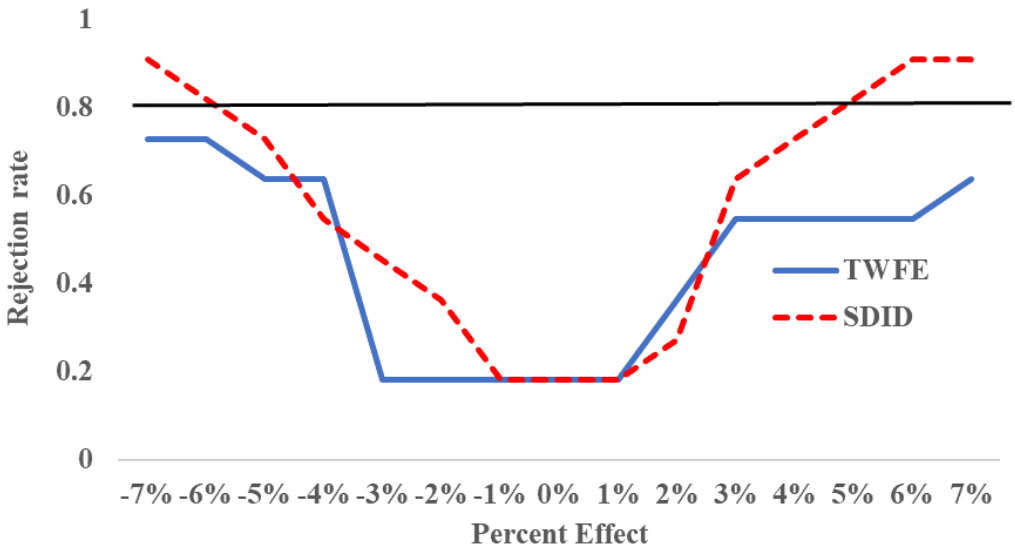
(a) Age 15-19

(b) Age 20-24

(c) Age 25-29

(d) Age 30

Figure B.8: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on 12 states excluding Texas at the end of every 5-year period from 2005-2019 by age group. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.



(a) Hispanic

Figure B.9: Synthetic difference-in-differences and two-way fixed effects power analysis rejection rates imposing effect sizes on 12 states excluding Texas at the end of every 5-year period from 2005-2019 for Hispanic women. Notes: We use fertility rates in each year-state as the outcome imposing effects from -7 to 7 percent of that year-state in the 12 randomly selected states in 2019, the last year of the power analysis. We count the number of rejected effects with t-statistics greater than 1.96 or less than -1.96 when the last year is considered treated.