The Econometrics Journal

Econometrics Journal (2022), volume **00**, pp. 1–23.
https://doi.org/10.1093/ectj/utac010

The triple difference estimator

ANDREAS OLDEN AND JARLE MØEN

Dept. of Business and Management Science, NHH Norwegian School of Economics, Hellevn. 30, N-5045 Bergen, Norway.
Email: andreasolden@gmail.com, jarle.moen@nhh.no

First version received: 14 May 2020; final version accepted: 10 May 2021.

Summary: Triple difference has become a widely used estimator in empirical work. A close reading of articles in top economics journals reveals that the use of the estimator to a large extent rests on intuition. The identifying assumptions are neither formally derived nor generally agreed on. We give a complete presentation of the triple difference estimator, and show that even though the estimator can be computed as the difference between two difference-in-differences estimators, it does not require two parallel trend assumptions to have a causal interpretation. The reason is that the difference between two biased difference-in-differences estimators will be unbiased as long as the bias is the same in both estimators. This requires only one parallel trend assumption to hold.

Keywords: *DD*, *DDD*, *DID*, *DiDiD*, *difference-in-difference-in-differences*, *difference-in-differences*, *parallel trend assumption*, *triple difference*.

JEL Codes: *C10*, *C18*, *C21*.

1. INTRODUCTION

The triple difference estimator is widely used, either under the name 'triple difference' (TD) or the name 'difference-in-difference-in-differences' (DDD), or with minor variations of these spellings. Triple difference is an extension of double differences and was introduced by Gruber (1994). Even though Gruber's paper is well cited, very few modern users of triple difference credit him for his methodological contribution. One reason may be that the properties of the triple difference estimator are considered obvious. Another reason may be that triple difference was little more than a curiosity in the first ten years after Gruber's paper. On Google Scholar, the annual number of references to triple difference did not pass one hundred until year 2007. Since then, the use of the estimator has grown rapidly and reached 928 unique works referencing it in the year 2017.[1]

Looking only at the core economics journals *American Economic Review* (AER), *Journal of Political Economy* (JPE), and *Quarterly Journal of Economics* (QJE), we have found 32 articles using triple difference between 2010 and 2017, see Table A1 in Appendix A. A close reading of these articles reveals that the use of the triple difference estimator to a large extent rests on

[1] More details on the historical development of the use of the triple difference estimator can be found in the working paper version of Olden and Møen (2020, fig. 1). In the working paper, we also analyse naming conventions and suggest that there is a need to unify terminology. We recommend the terms 'triple difference' and 'difference-in-difference-in-differences'.

intuition. The identifying assumptions are neither formally derived nor generally agreed on. We fill this void in the literature and give a complete presentation of the triple difference estimator.

The triple difference estimator can be computed as the difference between two difference-in-differences estimators. Despite this, we show that the triple difference estimator does not require two parallel trend assumptions to have a causal interpretation. The intuition is that the difference between two biased difference-in-differences estimators will be unbiased as long as the bias is the same in both estimators. In that case, the bias will be differenced out when the triple difference is computed. This requires only one parallel trend assumption, in ratios, to hold. In fact, the sole purpose of subtracting the second difference-in-differences is to remove bias in the first. Gruber (1994) states the identification requirement verbally, but the result has not been fully formalised in the econometric literature, and it is overlooked in most of the recent applications.

The rest of the paper is organised as follows: Section 2 gives a short overview of the use of the triple difference estimator. Section 3 derives the triple difference estimator. Section 4 shows that the triple difference estimator can be viewed as the difference between two difference-in-differences estimators. Section 5 derives the identifying assumptions. Section 6 shows that the triple difference estimator can also be viewed as a difference-in-differences using a ratio between two outcome variables. Section 7 discusses some issues related to inference. Section 8 provides concluding remarks.

2. THE TRIPLE DIFFERENCE LITERATURE

The most authoritative and formal treatment of the triple difference estimator was for many years an unpublished NBER summer institute lecture note on difference-in-differences estimation by Imbens and Wooldridge (2007). In the introductory 'Review of Basic Methodology' chapter they included a brief exposition of the triple difference estimator.[2] The formula for the triple difference estimator is now available in two econometrics books by Frölich and Sperlich (2019, p. 242) and Wooldridge (2020, p. 436). We complement these recent books by providing a more detailed discussion of the estimator, and in particular by deriving the assumptions needed to identify a causal effect.[3]

Other authoritative sources have treated the topic only in passing. In their famous text book, *Mostly Harmless Econometrics*, Angrist and Pischke (2008, p. 242) write that 'A modification of the two-by-two DD setup with possibly improved control groups uses higher-order contrast to draw causal inference'. The authors then go on to explain the basic setup using Yelowitz (1995) as an example. They do not discuss or present the estimator, nor the identifying assumption. They simply conclude that 'This triple-difference model may generate a more convincing set of results than a traditional DD analysis'.

Lechner (2011, p. 3) follows a similar avenue in his survey *The Estimation of Causal Effects by Difference-in-Differences Methods*. He uses Yelowitz (1995) as an example of triple difference and states that 'the basic ideas of the approach of taking multiple differences are already apparent

[2] Imbens and Wooldridge (2007) start out with a setup that is identical to ours in all respects except notation (compare their equation 1.3 to our (3.1)) However, the estimator presented in their equation 1.4, contains an error as it lacks the last term in our (3.4). This was corrected already in the 2008-version of the lecture notes, but unfortunately, later versions have been less widely distributed.

[3] We are grateful to an anonymous referee for making us aware of the two recent books.

with two dimensions. Thus, we refrain from addressing these higher dimensions to keep the discussion as focused as possible.'

A look at Yelowitz (1995) reveals that he does not go into depth on the estimator and the identifying assumptions. Instead, he cites Gruber (1994) and Gruber and Poterba (1994). Gruber and Poterba (1994), however, refer back to Gruber (1994).

In his single-authored 1994 article, Gruber analyses the labour market effects of mandated maternity benefits. Gruber explains the setup as follows:

> I compare the treatment individuals in the experimental states to a set of control individuals in those same states and measure the change in the treatments' relative outcomes, relative to states that did not pass maternity mandates. The identifying assumption of this 'differences-in-differences-in-differences' (DDD) estimator are fairly weak: it simply requires that there be no contemporaneous shock that affects the relative outcomes of the treatment group in the same state-years as the law.

We have also looked at all articles applying triple difference (using one of the six most common ways of referencing the estimator) in *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics* between 2010 and 2017. As seen in Table A1 in Appendix A, we found a total of 32 articles, 16 articles in AER, five in JPE and 11 in QJE. Of these articles Muehlenbachs et al. (2015), Hornbeck (2010), and Shayo and Zussman (2011) show some version of the estimator itself, indicating that it is not entirely obvious. In a similar spirit, Walker (2013) shows the error term of the triple difference estimator and uses it for discussion of robustness. Only Nilsson (2017) cites Gruber (1994).

We will later show formally that a parallel trend assumption very similar to the difference-in-differences approach is needed for the estimated effect to have a causal interpretation. The parallel trend in DDD is, however, on a differential between two categories. In some applications this is stated verbally. Walker (2013, p. 1805), e.g., writes that '[t]he identifying assumption in this class of models is that there are no other factors generating a difference in differential trends between production decisions in regulated and unregulated manufacturing firms.'[4]

Most of the other 32 top journal articles present some intuition of what the estimator is robust against, but otherwise the information presented varies considerably. Only a few of the authors discuss a common trend or parallel trend assumption, and as the triple difference is based on a strong parallel trend assumption, it is also disturbing to see that a large part of the articles do not include unconditional plots of the outcome series that they are studying. This makes it impossible to visually assess potential trends.

In Tables A2a and A2b in Appendix A, we present the 50 most cited articles referencing the estimator, numbered and ordered by number of citations. There has been almost 5,000 papers referencing the estimator since 1994, and it is natural to think that some of the most cited triple difference articles are methodological or represent early use of the methodology. Seven of the 50 most cited articles list Gruber as a co-author.[5] Six articles are covered in the review of articles in AER/QJE/JPE.[6] Among the rest, seven have methodological-sounding names.[7] A close reading of

[4] Some other articles in our sample have similar formulations. Hoynes et al. (2016, p. 925–6) write that '[i]n this triple-difference model, the maintained assumption is that there are no differential trends for high participation versus low participation groups within early versus late implementing counties'. Deschênes et al. (2017, p. 2970) state that '[o]ur identifying assumption is that such policies did not change differentially in NBP versus non-NBP states, in winter versus summer, over this period'. Finally, Kleven et al. (2013, p. 1908) write that '[i]n that case, the identifying assumption would be that there is no contemporaneous change in the differential trend between Spain and the synthetic control country'.

[5] These are the articles 4, 9, 17, 25, 31, 34, and 39, in which 4 is Gruber (1994) and 31 is Gruber and Poterba (1994). Note also that number 30 is Yelowitz (1995).

[6] These are the articles 7, 11, 21, 35, 42, and 46.

[7] These are the articles 1, 5, 6, 10, 12, 24, and 40. Note that number 24 is Lechner (2011) which is covered previously.

the articles with methodological-sounding names reveals that they do not give a formal exposition of the triple difference estimator, nor its identifying assumption. However, Ravallion (2007) cites Ravallion et al. (2005) which shows a very special case of the triple difference estimator and the identifying assumptions for that special case.[8]

3. THE TRIPLE DIFFERENCE ESTIMATOR

For the sake of exposition let us assume that we are talking about two American states, and that the treatment state (T) introduces a health-care measure, while the control state (C) does not. Further, the population of the states can be subdivided into two groups, group A and group B. The health-care measure we intend to study is only introduced to group B, i.e., group B is the group that can *B*enefit from the measure. Finally, there are two time periods, namely pre- and post-implementation of the health-care measure.

To establish a counterfactual it might seem convenient to compare group A and group B within the treatment state. This will not be valid if the health-care reform has within-state spillovers from group B to group A. Another option is to compare group B in the treatment state with group B in the control state. This will not be valid if different states have different economic conditions, so that group B in the treatment state would have trended differently from group B in the control state, regardless of the health-care measure. However, we may reasonably assume that the general economic differences will not affect the relative outcomes of group A and group B. In that case, we can use the relative difference to estimate what would have happened to the relative outcomes of group A and group B in the treatment state in the absence of treatment.

Equation (3.1) is a basic triple difference specification in accordance with the above exposition. All variables in this basic setup are dummy variables.

$$\begin{aligned} Y_{sit} = {} & \beta_0 + \beta_1 T + \beta_2 B + \beta_3 Post + \beta_4 T \times B + \beta_5 T \times Post + \beta_6 B \times Post \\ & + \beta_7 T \times B \times Post + \epsilon_{sit}. \end{aligned} \tag{3.1}$$

The conditional mean function of (3.1) is $E[Y_{sit}|T, B, Post]$, which can take on eight values. Since the model has eight values and eight coefficients, the model is saturated (Angrist and Pischke, 2008). Under standard OLS assumptions and an additive effect, we can use $E[\epsilon_{sit}|T, B, Post] = 0$ to show the eight expected values as in (3.2).

$$\begin{aligned} E[Y|T = 0, B = 0, Post = 0] &= \beta_0 \\ E[Y|T = 1, B = 0, Post = 0] &= \beta_0 + \beta_1 \\ E[Y|T = 0, B = 1, Post = 0] &= \beta_0 + \beta_2 \\ E[Y|T = 0, B = 0, Post = 1] &= \beta_0 + \beta_3 \\ E[Y|T = 1, B = 1, Post = 0] &= \beta_0 + \beta_1 + \beta_2 + \beta_4 \\ E[Y|T = 1, B = 0, Post = 1] &= \beta_0 + \beta_1 + \beta_3 + \beta_5 \\ E[Y|T = 0, B = 1, Post = 1] &= \beta_0 + \beta_2 + \beta_3 + \beta_6 \\ E[Y|T = 1, B = 1, Post = 1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7. \end{aligned} \tag{3.2}$$

[8] This scenario does not have pre-periods, only post-periods, and two treatment groups that are treated with differential intensity. This requires a set of identifying assumptions that in general are not needed in the triple difference estimator.

Starting at the top of equation set (3.2), we can solve for the $\beta's$.

$$
\begin{aligned}
\beta_0 &= E[Y|T=0, B=0, Post=0] \\
\beta_1 &= E[Y|T=1, B=0, Post=0] - E[Y|T=0, B=0, Post=0] \\
\beta_2 &= E[Y|T=0, B=1, Post=0] - E[Y|T=0, B=0, Post=0] \\
\beta_3 &= E[Y|T=0, B=0, Post=1] - E[Y|T=0, B=0, Post=0] \\
\beta_4 &= E[Y|T=1, B=1, Post=0] + E[Y|T=0, B=0, Post=0] \\
&\quad - E[Y|T=1, B=0, Post=0] - E[Y|T=0, B=1, Post=0] \\
\beta_5 &= E[Y|T=1, B=0, Post=1] + E[Y|T=0, B=0, Post=0] \\
&\quad - E[Y|T=1, B=0, Post=0] - E[Y|T=0, B=0, Post=1] \\
\beta_6 &= E[Y|T=0, B=1, Post=1] + E[Y|T=0, B=0, Post=0] \\
&\quad - E[Y|T=0, B=1, Post=0] - E[Y|T=0, B=0, Post=1] \\
\beta_7 &= \big(E[Y|T=1, B=1, Post=1] - E[Y|T=1, B=1, Post=0]\big) \\
&\quad - \big(E[Y|T=1, B=0, Post=1] - E[Y|T=1, B=0, Post=0]\big) \\
&\quad - \big(E[Y|T=0, B=1, Post=1] - E[Y|T=0, B=1, Post=0]\big) \\
&\quad + \big(E[Y|T=0, B=0, Post=1] - E[Y|T=0, B=0, Post=0]\big). \qquad (3.3)
\end{aligned}
$$

By rearranging the expression for β_7 and substituting the expected values with their sample equivalents (the mean values), we get (3.4). This is the triple difference estimator for the effect of the treatment for group B.

$$
\begin{aligned}
\hat{\beta_7} &= [(\bar{Y}_{T=1,B=1,Post=1} - \bar{Y}_{T=1,B=1,Post=0}) - (\bar{Y}_{T=0,B=1,Post=1} - \bar{Y}_{T=0,B=1,Post=0})] \\
&\quad - [(\bar{Y}_{T=1,B=0,Post=1} - \bar{Y}_{T=1,B=0,Post=0}) - (\bar{Y}_{T=0,B=0,Post=1} - \bar{Y}_{T=0,B=0,Post=0})]. \qquad (3.4)
\end{aligned}
$$

For simplicity, we have not included control variables in the equations above. Adding control variables is common and simple when using the regression formulation of the triple difference model in (3.1). The benefits are two-fold. First, control variables with substantial explanatory power will reduce the residual variance, and thereby increase the precision of the causal effect of interest. Second, including control variables can account for compositional differences between groups and make the parallel trend assumption needed for identification more credible. Put differently, including control variables can mitigate selection problems if there is some selection into the treatment state and group that is based on observable characteristics. We derive the identifying assumption for the case without control variables in Section 5.

4. THE DIFFERENCE BETWEEN TWO DIFFERENCE-IN-DIFFERENCES

The classical difference-in-differences estimator is presented in (4.1).

$$
\hat{\delta} = [(\bar{Y}_{T=1,Post=1} - \bar{Y}_{T=1,Post=0}) - (\bar{Y}_{T=0,Post=1} - \bar{Y}_{T=0,Post=0})]. \qquad (4.1)
$$

Clearly, the triple difference estimator of (3.4) is equivalent to the difference between two difference-in-differences. The first difference-in-differences is for group B, and is given by the first square brackets, while the second difference-in-differences is for group A, given by the second square brackets. It is also worth mentioning that due to the additive nature of the triple difference estimator of (3.4), we could alternatively have presented it as a difference-in-differences for the treatment state, comparing the eligible group B and group A, minus a difference-in-differences in the control state, comparing group B and group A there. Mathematically this is equivalent, though when thinking about a specific application, one is often preferred over the other.

5. IDENTIFYING ASSUMPTIONS

The triple difference estimator requires a parallel trend assumption for the estimated effect to have a causal interpretation. Even though the triple difference is the difference between two difference-in-differences, it does not need two parallel trend assumptions. Rather, it requires the relative outcome of group B and group A in the treatment state to trend in the same way as the relative outcome of group B and group A in the control state in the absence of treatment.[9] To see this, first take the β_7 in (3.3) and rearrange it to create (5.1).

$$\begin{aligned}\beta_7 = &\Big[\Big(E[Y|T=1,B=1,Post=1]-E[Y|T=1,B=1,Post=0]\Big)\\ &-\Big(E[Y|T=1,B=0,Post=1]-E[Y|T=1,B=0,Post=0]\Big)\Big]\\ &-\Big[\Big(E[Y|T=0,B=1,Post=1]-E[Y|T=0,B=1,Post=0]\Big)\\ &-\Big(E[Y|T=0,B=0,Post=1]-E[Y|T=0,B=0,Post=0]\Big)\Big]. \qquad (5.1)\end{aligned}$$

Now, introduce the potential outcomes framework (see, for instance, Angrist and Pischke, 2008). In this framework $E[Y_{1,sit}]$ is the expected outcome of a state, group, and time if treated, while $E[Y_{0,sit}]$ is the expected outcome of a state, group, and time if not treated. Potential outcomes mean that we either observe $\overline{Y}_{1,sit}$ or $\overline{Y}_{0,sit}$, but never both. Expressions like $E[Y_{0,T=1,B=1,Post=1}]$ are the expectation of non-observed potential outcomes; in our case the outcome of group B in the treatment state (T), in the treatment period (*Post*), had it not been treated.

We can use the potential outcome framework to define δ, the true causal effect of treatment in the treatment state (T), on the treatment group B, in the treatment period (*Post*) as:

$$\delta = E[Y_1 - Y_0|T=1,B=1,Post=1]. \qquad (5.2)$$

Equation (5.2) states that the true treatment effect is the difference between the outcome of state T, group B in period 2 as treated, and the outcome of state T, group B in period 2, had it not been treated.

[9] We phrase the discussion here in terms of trends, but, as mentioned in the introduction, one can also think of triple difference as a way to remove a potential bias in an ordinary difference estimator. This requires that the two DD-estimators used have the same bias. In fact, even the ordinary difference-in-differences estimator can in general terms be thought of as a way to remove bias rather than time trends. The parallel trend assumption is therefore sometimes referred to as a 'bias stability' assumption, see, e.g., Frölich and Sperlich (2019, p. 230).

Note that (5.2) is the *average treatment effect on the treated*, often called ATET, ATT or TOT, see, e.g., Angrist and Pischke (2008, ch. 3). Under the parallel trend assumption this is what is identified. This can be seen from the conditioning on $T = 1$ in the definition of the true causal effect, δ. With heterogeneous treatment effects, the population wide, unconditional, average treatment effect (ATE) is not identified. In the DD case, this has previously been pointed out by Frölich and Sperlich (2019, p. 228). They explain this by the fact that treatment effect estimation using the difference-in-differences estimator is a prediction problem where outcomes observed before the treatment started are used to predict the potential non-treatment outcome. With heterogeneous treatment effects, however, the natural experiments used for difference-in-differences estimation do not necessarily contain any information to predict the potential treatment outcome for the control group. This reasoning also applies to triple difference estimation where there are three non-predictable, counter factual, treatment outcomes, $E[Y_{1,T=0,B=1,Post=1}]$, $E[Y_{1,T=0,B=0,Post=1}]$, and $E[Y_{1,T=1,B=0,Post=1}]$.

We are now ready to derive the parallel trend assumption that identifies δ. Doing so, we rewrite (5.1) using the notation from the potential outcome framework.

$$\begin{aligned}\beta_7 = &\Big[\Big(E[Y_1|T=1,B=1,Post=1] - E[Y_0|T=1,B=1,Post=0]\Big)\\ &- \Big(E[Y_0|T=1,B=0,Post=1] - E[Y_0|T=1,B=0,Post=0]\Big)\Big]\\ &- \Big[\Big(E[Y_0|T=0,B=1,Post=1] - E[Y_0|T=0,B=1,Post=0]\Big)\\ &- \Big(E[Y_0|T=0,B=0,Post=1] - E[Y_0|T=0,B=0,Post=0]\Big)\Big]. \qquad (5.3)\end{aligned}$$

For β_7 to equal δ, we need the differential in the outcomes of group A and group B in the treatment state to trend similarly to the differential in the outcomes of group A and group B in the control state, in the absence of treatment. This is the parallel trend assumption. A formal exposition of this statement is given in (5.4). The first line is the change between the two periods in the outcomes of group B in the treatment state had it not been treated. The second line is the same change for group A. The difference between these two expressions is equated with an expression that is equivalent, except that it gives realised outcomes in the control state.[10]

$$\begin{aligned}&\Big(E[Y_0|T=1,B=1,Post=1] - E[Y_0|T=1,B=1,Post=0]\Big)\\ &- \Big(E[Y_0|T=1,B=0,Post=1] - E[Y_0|T=1,B=0,Post=0]\Big)\\ &\qquad\qquad =\\ &\Big(E[Y_0|T=0,B=1,Post=1] - E[Y_0|T=0,B=1,Post=0]\Big)\\ &- \Big(E[Y_0|T=0,B=0,Post=1] - E[Y_0|T=0,B=0,Post=0]\Big). \qquad (5.4)\end{aligned}$$

[10] See Frölich and Sperlich (2019, p. 244) for a different formulation given in the context of DDD used on a three period, two group setup. The DD parallel trend assumption then translates into what they call a 'parallel growth' or 'common acceleration' assumption.

To show that this parallel trend assumption identifies δ, the causal effect, we can substitute (5.4) into (5.3).

$$\begin{aligned}\beta_7 = &\Big[\Big(E[Y_1|T=1,B=1,Post=1] - E[Y_0|T=1,B=1,Post=0]\Big)\\ &- \Big(E[Y_0|T=1,B=0,Post=1] - E[Y_0|T=1,B=0,Post=0]\Big)\Big]\\ &- \Big[\Big(E[Y_0|T=1,B=1,Post=1] - E[Y_0|T=1,B=1,Post=0]\Big)\\ &- \Big(E[Y_0|T=1,B=0,Post=1] - E[Y_0|T=1,B=0,Post=0]\Big)\Big]. \qquad (5.5)\end{aligned}$$

Rearranging and rewriting (5.5) we get

$$\begin{aligned}\beta_7 = &E[Y_1 - Y_0|T=1,B=1,Post=1]\\ &+ E[Y_0|T=1,B=0,Post=1] - E[Y_0|T=1,B=0,Post=1]\\ &+ E[Y_0|T=1,B=1,Post=0] - E[Y_0|T=1,B=1,Post=0]\\ &+ E[Y_0|T=1,B=0,Post=0] - E[Y_0|T=1,B=0,Post=0]. \qquad (5.6)\end{aligned}$$

By cancelling out the redundant terms of (5.6) we arrive at (5.7)

$$\beta_7 = E[Y_1 - Y_0|T=1,B=1,Post=1] = \delta \quad q.e.d. \qquad (5.7)$$

6. TRIPLE DIFFERENCE AS DIFFERENCE-IN-DIFFERENCES

Take the difference-in-differences estimator of (4.1) and define the outcome variable, $\bar{Y}$, as:

$$\bar{Y}_{ij} = \bar{Y}_{aij} - \bar{Y}_{bij}. \qquad (6.1)$$

Substituting (6.1) back into (4.1) gives us (6.2)

$$\begin{aligned}\hat{\delta} = &[(\bar{Y}_{a,pre,treat} - \bar{Y}_{b,pre,treat}) - (\bar{Y}_{a,post,treat} - \bar{Y}_{b,post,treat})]\\ &- [(\bar{Y}_{a,pre,cont} - \bar{Y}_{b,pre,cont}) - (\bar{Y}_{a,post,cont} - \bar{Y}_{b,post,cont})]\\ = &\hat{\delta}_{triple}. \qquad (6.2)\end{aligned}$$

This shows clearly that a basic difference-in-differences with a differential as the outcome and a symmetric structure is a triple difference, and the other way around. This implies that all procedures for difference-in-differences can be applied to a transformed triple difference. For instance, standard robustness checks for difference-in-differences can be applied, see, for example, Angrist and Pischke (2008). Also, semi-parametric versions of the difference-in-differences estimator are available, as in Abadie (2005), as well as non-linear models as in Athey and Imbens (2006), which can be directly applied to the transformed problem. Among the generalisation of the simple difference-in-differences estimator, Callaway and SantAnna (2020) provides an appropriate estimators for cases with many time periods, including when the parallel trend assumption holds only conditional on covariates. They also give an up to date literature review.

Finally, knowing that difference-in-differences models struggle with standard errors when there are few clusters, as documented by Bertrand et al. (2004), this will apply to the transformed triple difference as well as to the triple difference estimator. We return to this next.

7. INFERENCE

In the case of the difference-in-differences estimator, Bertrand et al. (2004) show how the estimator is prone to over-rejection, i.e., finding false positives. This is due to serial correlation and intra-group correlation. This can be addressed by using cluster robust standard errors, which are based on asymptotic properties in the group dimension. However, it is common to have a limited number groups or treatment groups, violating the assumptions. For a fairly recent and extensive exposition of the issues in the difference-in-differences estimator, see Cameron and Miller (2015).

It is unclear to what extent this generalises to the triple difference estimator as we include additional groups, correlation structures, and explicitly try to model them. Also, we increase the number of observations and the complexity of the model. To answer these questions, we turn to a procedure from Bertrand et al. (2004) running a simulation study on data from the Current Population Survey (CPS) in which we vary the number of treated clusters.[11] We compare the difference-in-differences estimator with the triple difference estimator, both for individual level data and for state-year-gender aggregated data. Further, we include the triple difference estimated as difference-in-differences on a ratio (see Section 6). The full results are presented in Appendix B.

When it comes to false positives, or over-rejection, we find that the difference-in-differences and the triple difference show similar patterns of over-rejection with clustered standard errors. However, the triple difference shows greater power to detect true (simulated) effects. Aggregation does not solve the issues of over-rejection and comes at a cost with respect to power. Further, the triple difference as difference-in-differences and the full triple difference performs almost identical. Researchers should know that there is little to lose, and some to gain, by using the triple difference relative to difference-in-differences, but also realise that when there are few clusters, or few treated clusters, both will have severe issues of over-rejection.

8. CONCLUDING REMARKS

In this paper we document the rise of the triple difference estimator. The use of the estimator has grown exponentially, yet it lacks formal derivation and is often carelessly applied in the literature, for instance, by largely ignoring its parallel trend assumption and by omitting unconditional plots, making model validation difficult.

Our main contribution is to show that the triple difference estimator does not require two parallel trend assumptions to have a causal interpretation, even though it can be computed as the difference between two difference-in-differences estimators. We also show that the triple difference parallel

[11] We draw n placebo treatment states out of 51, draw a year from a uniform distribution over 1985–1995 which serves as a treatment year, estimate different models, and reiterate the process 10,000 times, considering rejection rates, i.e., how often we find a significant effect.

trend assumption is equivalent to the parallel trend assumption in a difference-in-differences model based on ratios.

When choosing between a triple difference and a difference-in-differences on a ratio-variable, there are several things to consider. The difference-in-differences estimator is much better understood, and there is a large literature that addresses the estimator and its shortcomings. However, it comes at the cost of degrees of freedom and provides less information than the triple difference. The triple difference will, for instance, provide an estimate of spillover effects, i.e., β_5 in (3.1), which is the effect on the non-treated in the treatment state in the treatment period. This information is lost in the difference-in-differences estimator.

The triple difference estimator is often used as a heterogeneity test or as a robustness check. When comparing it with a standard difference-in-differences, Berck and Villas-Boas (2016) show conditions for when the triple difference estimator reduces bias relative to a difference-in-differences approach in the presence of omitted variable bias.

Finally, our reading of the literature points to some other key issues that demand awareness. Many of the articles spend considerable time on control variables in which case one should be specific on whether the inclusion is to absorb variance and increase precision, or if the parallel trend assumption holds only conditional on some covariates. Note that in the case of time-invariant state-level variables, they will be differenced out, easily shown by deducting any mean from the estimator. Time-varying, state-level variables, however, is a likely source of bias and should be explicitly dealt with when evaluating the parallel trend assumption, or be dealt with in a more complex framework as touched upon in Section 6.

In the literature, much less time is spent discussing functional form issues than control variables. This is unfortunate. Both the difference-in-differences and the triple difference estimator relies on a parallel trend assumption, and hence the functional form is identifying. In the triple difference estimator, we make an assumption on how the outcomes of two groups co-move relative to the co-movement in two other groups in the control state. Both a ratio and its log-transformed counterpart can be a natural choice of functional form, depending on the situation. However, this requires thought. When the parallel trend assumption holds in logs it will not hold in levels, and vice versa, see Angrist and Pischke (2008, p. 230) and Frölich and Sperlich (2019, p. 228).[12]

ACKNOWLEDGEMENTS

This paper is a methodological companion paper to Olden (2018). We thank two anonymous referees for very helpful and valuable comments. We are also grateful to Erik Øiolf Sørensen and Håkon Otneim for useful discussions and comments. The paper is partly financed by the Research Council of Norway, Grant No. 267423.

[12] Unfortunately, what functional form to choose, seldom finds it answer in economic theory or statistics. For a discussion of these topics in the case of the difference-in-differences estimator, we recommend Kahn-Lang and Lang (2020). The recommendations of Kahn-Lang and Lang (2020) is equally applicable to the triple difference estimator and includes addressing why there is level differences to begin with, explicitly justifying the parallel trend assumption and noting that pre-treatment trends is indicative, but not necessary, nor sufficient for the parallel trend assumption to hold. However, in the case of the triple difference, initial level differences in the difference-in-differences might be a reason why we want to use triple difference. The general advice to reflect on level differences still stands.

REFERENCES

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1–19.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics 49*(4), 431–4.

Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*(2), 431–97.

Berck, P. and S. B. Villas-Boas (2016). A note on the triple difference in economic models. *Applied Economics Letters 23*(4), 239–42.

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm. CREA Discussion paper 13, University of Luxembourg.

Bertrand, M., E. Duflo and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*(1), 249–75.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–30.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–72.

Conley, T. G. and C. R. Taber (2011). Inference with 'difference in differences' with a small number of policy changes. *Review of Economics and Statistics 93*(1), 113–25.

Deschênes, O., M. Greenstone and J. S. Shapiro (2017). Defensive investments and the demand for air quality: Evidence from the NOx budget program. *American Economic Review 107*(10), 2958–89.

Ferman, B. and C. Pinto (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *Review of Economics and Statistics 101*(3), 452–67.

Frölich, M. and S. Sperlich (2019). *Impact Evaluation Treatment Effects and Causal Analysis*. Cambridge: Cambridge University Press.

Gruber, J. (1994). The incidence of mandated maternity benefits. *American Economic Review 84*(3), 622–41.

Gruber, J. and J. Poterba (1994). Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *Quarterly Journal of Economics 109*(3), 701–33.

Hornbeck, R. (2010). Barbed wire: Property rights and agricultural development. *Quarterly Journal of Economics 125*(2), 767–810.

Hoynes, H., D. W. Schanzenbach and D. Almond (2016). Long-run impacts of childhood access to the safety net. *American Economic Review 106*(4), 903–34.

Imbens, G. W. and J. M. Wooldridge (2007). What's new in econometrics? Difference-in-differences estimation. Lecture 10 presented at NBER Summer Institute, NBER.

Kahn-Lang, A. and K. Lang (2020). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business and Economic Statistics 38*(3), 613–20.

Kleven, H. J., C. Landais and E. Saez (2013). Taxation and international migration of superstars: Evidence from the European football market. *American Economic Review 103*(5), 1892–924.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics 4*(3), 165–224.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

MacKinnon, J. G. and M. D. Webb (2018). The wild bootstrap for few (treated) clusters. *Econometrics Journal 21*(2), 114–35.

MacKinnon, J. G. and M. D. Webb (2020). Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics 218*(2), 435–50.

Muehlenbachs, L., E. Spiller and C. Timmins (2015). The housing market impacts of shale gas development. *American Economic Review 105*(12), 3633–59.

National Bureau of Economic Research (1979–1999). Current population survey merged outgoing rotation groups repository. https://data.nber.org/morg/annual/ (accessed: 19 November 2020).

Nilsson, J. P. (2017). Alcohol availability, prenatal conditions, and long-term economic outcomes. *Journal of Political Economy 125*(4), 1149–207.

Olden, A. (2018). What do you buy when no one's watching? The effect of self-service checkouts on the composition of sales in retail. Discussion paper FOR 3/18, Norwegian School of Economics, Norway.

Olden, A. and J. Møen (2020). The triple difference estimator. Discussion paper FOR 1/20, Norwegian School of Economics, Norway.

Ravallion, M. (2007). Evaluating anti-poverty programs. *Handbook of Development Economics 4*, 3787–846.

Ravallion, M., E. Galasso, T. Lazo and E. Philipp (2005). What can ex-participants reveal about a program's impact? *Journal of Human Resources 40*(1), 208–30.

Shayo, M. and A. Zussman (2011). Judicial ingroup bias in the shadow of terrorism. *Quarterly Journal of Economics 126*(3), 1447–84.

Walker, W. R. (2013). The transitional costs of sectoral reallocation: Evidence from the clean air act and the workforce. *Quarterly Journal of Economics 128*(4), 1787–835.

White, H. (1984). *Asymptotic Theory for Econometricians*. San Diego, CA: Academic Press.

Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach*. 7th edn. Boston, MA: Cengage.

Yelowitz, A. S. (1995). The medicaid notch, labor supply, and welfare participation: Evidence from eligibility expansions. *Quarterly Journal of Economics 110*(4), 909–39.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication Package

Co-editor Petra Todd handled this manuscript.

APPENDIX A: TABLES

Table A1. Use of triple difference estimation in AER, JPE and QJE from 2010 to 2017.

Cites	Authors	Title	Year	Source
829	Mian, Sufi	House prices, home equity-based borrowing, and the US household leverage crisis	2011	AER
103	Moser, Voena	Compulsory licensing: Evidence from the trading with the enemy act	2012	AER
293	Hornbeck	The enduring impact of the American Dust Bowl: Short-and long-run adjustments to environmental catastrophe	2012	AER
146	Simcoe	Standard setting committees: Consensus governance for shared technology platforms	2012	AER
243	Kleven, Landais, Saez	Taxation and international migration of superstars: Evidence from the European football market	2013	AER
320	Busso, Gregory, Kline	Assessing the incidence and efficiency of a prominent place based policy	2013	AER
57	Aaronson, Lange, Mazumder	Fertility transitions along the extensive and intensive margins	2014	AER
129	Yagan	Capital tax reform and the real economy: The effects of the 2003 dividend tax cut	2015	AER
90	Casey	Crossing party lines: The effects of information on redistributive politics	2015	AER
212	Muehlenbachs, Spiller, Timmins	The housing market impacts of shale gas development	2015	AER
291	Hoynes, Schanzenbach, Almond	Long-run impacts of childhood access to the safety net	2016	AER
440	Pierce, Schott	The surprisingly swift decline of US manufacturing employment	2016	AER
37	Duggan, Garthwaite, Goyal	The market impacts of pharmaceutical product patents in developing countries: Evidence from India	2016	AER
65	Egan, Hortasu, Matvos	Deposit competition and financial fragility: Evidence from the us banking sector	2017	AER
30	Deschênes, Greenstone, Shapiro	Defensive investments and the demand for air quality: Evidence from the NOx budget program	2017	AER
122	Besley, Folke, Persson, Rickne	Gender quotas and the crisis of the mediocre man: Theory and evidence from Sweden	2017	AER
79	Aaronson, Mazumder	The impact of Rosenwald schools on black achievement	2011	JPE

Table A1. Continued

Cites	Authors	Title	Year	Source
50	Autor, Palmer, Pathak	Housing market spillovers: Evidence from the end of rent control in Cambridge, Massachusetts	2014	JPE
163	Carneiro, Lken, Salvanes	A flying start? Maternity leave benefits and long-run outcomes of children	2015	JPE
37	Casas-Arce, Saiz	Women and power: Unpopular, unwilling, or held back?	2015	JPE
47	Nilsson	Alcohol availability, prenatal conditions, and long-term economic outcomes	2017	JPE
143	Hornbeck	Barbed wire: Property rights and agricultural development	2010	QJE
179	Shayo, Zussman	Judicial ingroup bias in the shadow of terrorism	2011	QJE
772	Ahern, Dittmar	The changing of the boards: The impact on firm valuation of mandated female board representation	2012	QJE
73	Cascio, Washington	Valuing the vote: The redistribution of voting rights and state funds following the voting rights act of 1965	2013	QJE
150	Walker	The transitional costs of sectoral reallocation: Evidence from the clean air act and the workforce	2013	QJE
155	Garthwaite, Gross, Notowidigdo	Public health insurance, labor supply, and employment lock	2014	QJE
52	Casaburi, Troiano	Ghost-house busters: The electoral response to a large antitax evasion program	2015	QJE
16	Agan, Starr	Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment	2017	QJE
25	Alsan, Wanamaker	Tuskegee and the health of black men	2017	QJE
44	Bandiera, Burgess, Das, Gulesci, Rasul, Sulaiman	Labor markets and poverty in village economies	2017	QJE
20	Larcom, Rauch, Willems	The benefits of forced experimentation: striking evidence from the London underground network	2017	QJE

Table A2a. Top 50 most cited articles referencing triple difference.

	Cites	Authors	Title	Year	Source
1	7,550	Bertrand, Duflo, Mullainathan	How much should we trust differences-in-differences estimates?	2004	QJE
2	1,418	Verhoogen	Trade, quality upgrading, and wage inequality in the Mexican manufacturing sector	2008	QJE
3	1,306	Currie, Almond	Human capital development before age five	2011	HLE
4	1,177	Gruber	The incidence of mandated maternity benefits	1994	AER
5	989	Roberts, Whited	Endogeneity in empirical corporate finance	2013	HEF
6	943	Winship, Morgan	The estimation of causal effects from observational data	1999	ARS
7	824	Mian, Sufi	House prices, home equity-based borrowing, and the US household leverage crisis	2011	AER
8	809	Ruhm	The economic consequences of parental leave mandates: Lessons from Europe	1998	QJE
9	807	Currie, Gruber	Health insurance eligibility, utilization of medical care, and child health	1996	QJE
10	774	Ravallion	Evaluating anti-poverty programs	2007	HDE
11	763	Ahern, Dittmar	The changing of the boards: The impact on firm valuation of mandated female board representation	2012	QJE
12	697	Besley, Case	Unnatural experiments? Estimating the incidence of endogenous policies	2000	TEJ
13	690	Giroud, Mueller	Does corporate governance matter in competitive industries?	2010	JFE
14	659	Zervas, Proserpio, Byers	The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry	2017	JMR
15	648	Dynarski	Hope for whom? Financial aid for the middle class and its impact on college attendance	2000	NTJ
16	552	Costa, Kahn	Power couples: Changes in the locational choice of the college educated, 1940–1990	2000	QJE
17	526	Gruber	The incidence of payroll taxation: Evidence from Chile	1997	JLE
18	512	Purnanandam	Originate-to-distribute model and the subprime mortgage crisis	2010	RFS
19	505	Low	Managerial risk-taking behavior and equity-based compensation	2009	JFE
20	500	Puri, Rocholl, Steffen	Global retail lending in the aftermath of the US financial crisis: Distinguishing between supply and demand effects	2011	JFE
21	436	Pierce, Schott	The surprisingly swift decline of US manufacturing employment	2016	AER
22	388	Katz	Wage subsidies for the disadvantaged	1996	NBER
23	387	Sommers, Baicker, Epstein	Mortality and access to care among adults after state Medicaid expansions	2012	NEJM
24	384	Lechner	The estimation of causal effects by difference-in-difference methods	2011	FTE
25	377	Gruber	Disability insurance benefits and labor supply	2000	JPE
26	359	Goldfarb, Tucker	Privacy regulation and online advertising	2011	MS

Table A2b. Top 50 most cited articles referencing triple difference, continued.

	Cites	Authors	Title	Year	Source
27	354	Strauss, Thomas	Health over the life course	2007	HDE
28	353	Matsa, Miller	A female style in corporate leadership? Evidence from quotas	2013	AEJAE
29	350	Seru	Firm boundaries matter: Evidence from conglomerates and R&D activity	2014	JFE
30	343	Yelowitz	The Medicaid notch, labor supply, and welfare participation: Evidence from eligibility expansions	1995	QJE
31	333	Gruber, Poterba	Tax incentives and the decision to purchase health insurance: Evidence from the self-employed	1994	QJE
32	332	Milligan	Subsidizing the stork: New evidence on tax incentives and fertility	2005	RES
33	330	Currie	Inequality at birth: Some causes and consequences	2011	AER
34	328	Gruber, Madrian	Health insurance, labor supply, and job mobility: A critical review of the literature	2002	NBER
35	319	Busso, Gregory, Kline	Assessing the incidence and efficiency of a prominent place based policy	2013	AER
36	318	Eggleston, Ling, Qingyue, Lindelow, Wagstaff	Health service delivery in China: A literature review	2008	HE
37	314	Neumark, Zhang, Ciccarella	The effects of Wal-Mart on local labor markets	2008	JUE
38	311	Figlio	Testing, crime and punishment	2006	JPuE
39	309	Gruber	Health insurance and the labor market	2000	HHE
40	296	Nichols	Causal inference with observational data	2007	SJ
41	291	Jensen	Do private transfers 'displace' the benefits of public transfers? Evidence from South Africa	2004	JPuE

Table A2b. Continued

	Cites	Authors	Title	Year	Source
42	290	Hornbeck	The enduring impact of the American Dust Bowl: Short-and long-run adjustments to environmental catastrophe	2012	AER
43	287	Thomas, Beegle, Frankenberg, Sikoki, Strauss, Teruel	Education in a Crisis	2004	JDE
44	286	Rishika, Kumar, Janakiraman, Bezawada	The effect of customers' social media participation on customer visit frequency and profitability: An empirical investigation	2013	ISR
45	282	Clotfelter, Glennie, Ladd, Vigdor	Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina	2008	JPuE
46	281	Hoynes, Schanzenbach, Almond	Long-run impacts of childhood access to the safety net	2016	AER
47	277	Morse	Payday lenders: Heroes or villains?	2011	JFE
48	277	Cai, Chen, Fang	Observational learning: Evidence from a randomized natural field experiment	2009	AER
49	273	Acharya, Baghai, Subramanian	Labor laws and innovation	2013	JLaE
50	267	Weber	The effects of a natural gas boom on employment and income in Colorado, Texas, and Wyoming	2012	EE

Notes: This table is produced using the software Harzinger's Publish or Perish 6. A search using each of the six most common ways to reference the triple difference estimator is conducted from 1994 until October 2018, covering almost all results for the triple difference estimator. Each search is combined with the word economics. When removing books and duplicates, this yields 3,481 articles. The articles are sorted according to the number of citations, and the top 50 most cited articles are presented here. Full journal titles are found in Table A3.

Table A3. Title abbreviations for Tables A1–A2b.

Abbreviation	Full title
AEJAE	*American Economic Journal: Applied Economics*
AER	*The American Economic Review*
ARS	*Annual Review of Sociology*
EE	*Energy Economics*
FTE	*Foundations and Trends® in Econometrics*
HDE	*Handbook of Development Economics*
HE	*Health Economics*
HEF	*Handbook of the Economics of Finance*
HHE	*Handbook of HE*
HLE	*Handbook of Labor Economics*
ISR	*Information Systems Research*
JDE	*Journal of Development Economics*
JFE	*Journal of Financial Economics*
JLaE	*Journal of Law and Economics*
JLE	*Journal of Labor Economics*
JMR	*Journal of Marketing Research*
JPE	*Journal of Political Economy*
JPuE	*Journal of Public Economics*
JUE	*Journal of Urban Economics*
MS	*Management Science*
NBER	NBER Working Paper Series
NEJM	*New England Journal of Medicine*
NTJ	*National Tax Journal*
QJE	*Quarterly Journal of Economics*
RES	*Review of Economics and Statistics*
RFS	*The Review of Financial Studies*
SJ	*Stata Journal*
TEJ	*The Economic Journal*

APPENDIX B: SIMULATIONS

The difference-in-differences and the triple difference estimators often have a group and a time structure, for instance, individual level data in different US states over time, with some states being treated. This structure introduces issues of serial correlation and intra-group (cluster) correlation, which can lead to biased standard errors and severe over-rejection of the null hypothesis of no effect, famously documented in the difference-in-differences case by Bertrand et al. (2004). Typically, cluster robust standard errors on the state level are used. These relies on asymptotic properties in the number of groups.[13] For a thorough overview on the issues and remedies, see Cameron and Miller (2015) and Angrist and Pischke (2008). It can be shown that the asymptotic properties applies to both the number of untreated and the number of treated clusters (see Conley and Taber, 2011).

While the issue of clustered errors is well studied in the difference-in-differences estimator, it is not obvious to what extent it carries over to the triple difference estimator. One reason to expect differences is that we introduce new contrast groups that might affect correlation structures, and we explicitly try to model sub-groups within the cluster (for instance gender in state). Moreover, the number of observations typically doubles and we increase the general complexity of the modelling approach. We will not give a full exposition of these issues in the triple difference estimator case, but we will make some points by comparing triple difference to difference-in-differences.

To aid intuition, consider the following stylized example. Some US states introduce a legal reform to affect the wage of women. Having data on wage from both before and after the reform for all states, it seems well-suited for a difference-in-differences approach. In this example, the states that introduce the legal reform are the treatment states, while the states that do not are the control states. The time period before the reform is the pre-period, while the time period after the reform is the treatment period. However, we might be worried that the states that introduced the legislation to impact the wage of women would have had higher growth rates to begin with, such that the comparison of women in the treatment state and women in the control states would be biased. While the treatment states might trend differentially from the control states regardless of treatment, we believe that this trend affects men and women similarly. Thus we consider a triple difference estimator for which we compare the relative wage of women and men in the treatment states to the relative wage of women and men in the control states, circumventing the bias from the difference-in-differences estimator.

If the assumptions hold and the right functional form is chosen, this strategy will get rid of the bias in the estimation. However, we are left with the question of standard errors. To shine some light on the issue, we use the procedure of Bertrand et al. (2004) and run simulations of placebo treatments while observing the rejection rates. The data used is the Current Population Survey in their fourth interview month, in the Merged Outgoing Rotation Group, from 1979 to 1999, see National Bureau of Economic Research (1979–1999). The survey contains individual level data from all US states and Washington DC. The data are freely available and commonly used.[14] The rejection rate is defined as the proportion of times the null hypothesis is rejected on a 5 percent significance level, i.e., the number of times we find a significant effect for the treatment variable. When there is no effect, this should be 5 percent of the times, i.e., the significance level or probability of false positives. Note that we are 'randomizing the treatment variable while keeping the set of outcomes fixed. In general, the distribution of the test statistic induced by such randomization is not a standard normal distribution and, therefore, the exact rejection rate we should expect is not known.' (Bertrand et al., 2004, p. 256). However, real data has its own advantages, and we also have the original article as a baseline. Furthermore, our comparisons are mainly to explore the relative performance of triple difference as compared to difference-in-differences, making the true rejection rate less important.

Keeping our example as close to Bertrand et al. (2004) as possible, we restrict the sample to participants between the ages of 25 and 50 with strictly positive earnings. This leaves about 1,000,000 observations.

[13] As developed by White (1984) with extensions by Liang and Zeger (1986) and Arellano (1987).

[14] Data accessed 19 November 2020 from https://www.nber.org/research/data/current-population-survey-cps-data-nber. Data and reproducible code is provided openly at https://github.com/andreasolden/simulate_triple_difference.

Bertrand et al. (2004) consider a difference-in-differences on women. We include men also, as they serve as a control group when adding the additional layer of the triple difference.[15] The procedure goes as follows:

(1) Draw n states randomly. These will serve as the placebo treatment states.
(2) Draw a year from a uniform distribution over 1985–1995. This year and all subsequent years will serve as the placebo treatment years.
(3) Esimate different models with different standard errors.
(4) Reiterate steps 1–3 10,000 times.
(5) Consider rejection rates, i.e., how often we find a significant effect.

We run five different regression models. Equation (B.1) is a difference-in-differences on females, as in Bertrand et al. (2004). Equation (B.2) is a triple difference for both sexes. Both are estimated on individual level data. Equation (B.3) is a difference-in-differences for females on data aggregated to the state-year-gender level. Equation (B.4) is a triple difference performed as difference-in-differences on relative outcomes for both sexes, as shown in Section 6. Equation (B.5) is a full triple difference for both sexes. The last two equations are also on data aggregated to the state-year-gender level. The motivation for the aggregation is that triple difference performed as difference-in-differences on relative outcomes is only possible for grouped data. Aggregation is sometimes also suggested as a way to circumvent intra-cluster correlation issues (Angrist and Pischke, 2008, p. 313). The outcome variable is always log-transformed weakly earnings, and s denotes state, i individual, and t time period.

$$logY_{sit}^{female} = \alpha_{states} + \gamma_{year} + \delta(T \times post) + \epsilon_{sit} \tag{B.1}$$

$$\begin{aligned} logY_{sit} = \alpha_{states} + \gamma_{year} + \beta_1(female) + \beta_2(T \times female) + \beta_3(post \times female) + \beta_4(T \times post) \\ + \delta(T \times post \times female) + \epsilon_{sit} \end{aligned} \tag{B.2}$$

$$log\bar{Y}_{st}^{females} = \alpha_{states} + \gamma_{year} + \delta(T \times post) + \epsilon_{st} \tag{B.3}$$

$$log(\bar{Y}_{st}^{females}/\bar{Y}_{st}^{males}) = \alpha_{states} + \gamma_{year} + \delta(T \times post) + \epsilon_{st} \tag{B.4}$$

$$\begin{aligned} log\bar{Y}_{st,gender} = \\ \alpha_{states} + \gamma_{year} + \beta_1(female) + \beta_2(T \times female) + \beta_3(post \times female) + \beta_4(T \times post) \\ + \delta(T \times post \times female) + \epsilon_{st,gender}. \end{aligned} \tag{B.5}$$

We repeat these estimations for 25, 5, 2, and 1 treated clusters, holding the total number of clusters constant at 51.[16] We also show results with different ways to estimate the standard errors. We use either uncorrected standard errors assuming independent and identically distributed errors (IID), White heteroscedasticity (HC1) robust standard errors (as Stata robust), or clustered standard errors on state-level (as Stata xtreg cluster). The implementation is done by the package Fixest in R (Bergé, 2018).[17] Finally, we simulate with

[15] We deviate from Bertrand et al. (2004) by running 10,000 iterations as opposed to 200–400. We do not include individual level controls for better comparisons between the simulated models, but we always include state and year fixed effects as well as a fixed effect of gender when applicable. Since the identification comes from group differences over time, this is unlikely to be important for our purposes.

[16] We deviate from Bertrand et al. (2004) who focus on the total number of clusters. However, as Conley and Taber (2011) point out, the asymptotics are for both treated and untreated groups, so we expect similar results as if we had just reduced the number of groups, holding the number of treated groups constant. We also expect this to be a more common issue as, even in the case of few clusters, the results will suffer from few treated clusters. For difference-in-differences, the scenario is covered in, for instance, Conley and Taber (2011), MacKinnon and Webb (2018), MacKinnon and Webb (2020), and Ferman and Pinto (2019).

[17] For more information see https://cran.r-project.org/package=fixest and https://cran.r-project.org/web/packages/fixest/vignettes/standard_errors.html.

Table B1. No effect: Rejection rates for individual level data models.

	Difference-in-differences			Triple difference		
	(1) Uncorrected	(2) White	(3) Cluster	(4) Uncorrected	(5) White	(6) Cluster
25	0.7058	0.7072	0.0711	0.3013	0.3021	0.0592
5	0.6757	0.6787	0.1530	0.3095	0.3106	0.1343
2	0.6187	0.6247	0.3421	0.3164	0.3192	0.3678
1	0.6015	0.6071	0.7408	0.2936	0.3096	0.8243
n	496,055	496,055	496,055	1,035,308	1,035,308	1,035,308

Notes: The table shows rejection rates for the treatment variable coefficient at a 5 percent significance level, over 10,000 simulations, individual level data, and a placebo treatment (no effect). Columns 1–3 is the difference-in-differences estimator, while columns 4–6 is the triple difference estimator. The columns differ in the standard errors that are used, where 1 and 4 makes no correction for correlation, 2 and 5 are White HC1 robust standard errors, while 4 and 6 are cluster robust standard errors. The row names indicate how many (placebo) treated clusters there were out of the total of 51.

Table B2. Rejection rates for individual level data models.

	No effect		2 percent effect		5 percent effect	
	(1) DD	(2) Triple	(3) DD	(4) Triple	(5) DD	(6) Triple
25	0.0711	0.0592	0.2148	0.5539	0.6941	0.9984
5	0.1530	0.1343	0.2910	0.3584	0.5761	0.8235
2	0.3421	0.3678	0.4213	0.4289	0.5803	0.7191
1	0.7408	0.8243	0.7830	0.7905	0.8391	0.9030
n	496,055	1,035,308	496,055	1,035,308	496,055	1,035,308

Notes: The table shows rejection rates for the treatment variable coefficient at a 5 percent significance level, over 10,000 simulations, individual level data, and a either a placebo treatment (columns 1 and 2), a 2 percent effect (columns 3 and 4), or a 5 percent effect (columns 5 and 6). Columns 1, 3, and 5 use the difference-in-differences estimator, while columns 2, 4, and 6 use the triple difference estimator. All standard errors are clustered at the state level. The row names indicate how many (placebo or real) treated clusters there were out of the total of 51.

Table B3. Rejection rates for aggregated models.

	No effect			2 percent effect			5 percent effect		
	(1) DD	(2) Triple as DD	(3) Triple	(4) DD	(5) Triple as DD	(6) Triple	(7) DD	(8) Triple as DD	(9) Triple
25	0.0479	0.0515	0.0517	0.1047	0.3101	0.3131	0.3958	0.9483	0.9488
5	0.1120	0.1235	0.1248	0.1562	0.2133	0.2148	0.3173	0.6303	0.6330
2	0.2894	0.3371	0.3391	0.3343	0.3714	0.3730	0.4118	0.5520	0.5526
1	0.7237	0.7929	0.7929	0.7490	0.8092	0.8092	0.7763	0.8495	0.8513
n	1,071	1,071	2,142	1,071	1,071	2,142	1,071	1,071	2,142

Notes: The table shows rejection rates for the treatment variable coefficient at a 5 percent significance level, over 10,000 simulations, state-year-gender aggregated data, and a either a placebo treatment (rows 1–3), a 2 percent effect (rows 4–6), or a 5 percent effect (rows 7–9). Rows 1, 4, and 7 use the triple difference estimator, rows 2, 5, and 8 use the triple difference as difference-in-differences (i.e., a difference-in-differences on a relative outcome), and rows 3, 6, and 9 is the full triple difference estimator. All standard errors are clustered at the state level. The column names indicate how many (placebo or real) treated clusters there were out of the total of 51.

a true effect of either 2 or 5 percent to get a sense of power or the ability to detect true effects. The effects are simulated by adding a fixed increase of 2 or 5 percent of the pre-treatment weakly earnings for women to the post-treatment outcomes for women in the treatment state, shifting the level, but not the trend. The results are shown in Tables B1, B2, and B3.

Results

Table B1 shows rejection rates for placebo treatments for individual level data. The estimators in use are the difference-in-differences of (B.1), in column 1–3, and the triple difference estimator of (B.2) in column 4–6. For each estimator, the rejection rates are either based on IID standard errors, robust standard errors, or clustered at the state level, in that order. Column 1 corresponds to the results from Bertrand et al. (2004), with rejection rates roughly between 60 and 70 percent, i.e., we find a significant effect 60 to 70 percent of the times in the case of IID standard errors and placebo treatment, which is severe over-rejection.[18] This is almost identical to the results with robust standard errors, as shown in column 2. Furthermore, the number of treated clusters, specified in the row names, has limited impact, going from rejection rates of about 70 percent to about 60 percent when reducing the number of treated clusters from 25 to 1. When clustering the standard errors, the rejection rate is 7 percent in the case of 25 treated clusters, which is close to what we would want, but it rises to 15 percent in the case of 5 treated clusters, 34 percent for 2 treated clusters, and 74 percent for 1 treated cluster, as expected when using standard errors based on cluster asymptotics.

Turning to the triple difference estimator, the results for IID and robust standard errors still over-reject, with rejection rates of about 30 percent, regardless of how many treated clusters there are, as shown in column 4 and 5. This is about half the rejection rate of the equivalent difference-in-differences, but still much higher than we would want. Finally, looking at column 6, the rejection rate is 6 percent for 25 treated clusters, which is close to ideal, 13 percent for 5 treated clusters, 37 percent for 2 treated clusters and 82 percent for 1 treated cluster, showing the same pattern as the difference-in-differences, however, mildly preferred for 5–25 treated clusters, but not for 1–2 treated clusters. The differences are, however, marginal. As for clustered errors, there seems to be little gain, nor loss, from moving from a difference-in-differences to a triple difference estimator, in terms of false positives.

Table B2 compares three scenarios, no effect (placebo treatments) as above, a simulated 2 percent true effect, and a simulated 5 percent true effect.[19] Each scenario contains both a difference-in-differences estimator and a triple difference estimator. All specifications are with individual level data, standard errors clustered at the state level, and either 25, 5, 2, or 1 treated clusters.

With many (25) treated clusters, the triple difference estimator shows signs of having more power than the difference-in-differences estimator, providing rejection rates of 55 percent to 21 percent in the case of a 2 percent effect, and 99.8 percent to 69 percent in the case of a 5 percent effect. When reducing the number of treated clusters to 5, the triple difference still has more power, with 36 percent to 29 percent for a 2 percent effect, and 82 percent to 57 percent for a 5 percent effect, which is better, but with smaller margins. With even fewer treated clusters the differences become slighter, and it is worth remembering that we get high rejection rates, even in the absence of treatment, when we have few treated clusters.

Table B3 are all performed on data aggregated to the state-year-gender level, see (B.3)–(B.5), and clustered standard errors. Aggregation to avoid intra-cluster correlation is common and suggested, for instance, by Angrist and Pischke (2008, p. 313). However, in the case of the (full) triple difference, we cannot aggregate to state-year, but have to aggregate to state-year-gender, leaving two observations per state per year, not fully avoiding the intra-cluster correlation structure. However, in the special case of (B.4), which is also shown (in levels) in Section 6, we estimate the triple difference as a difference-in-differences on a relative outcome, preserving the one observation per state per year structure.

There are some notable patterns. In the case of no effect and 25 treated clusters, the rejection rates differ only marginally, and are similar to the individual level estimations. As we decrease the number of

[18] Strictly speaking, column 1, row 1 is the same specification as row 1 in table II in (Bertrand et al., 2004, p. 257), except for the choice of covariates and the number of simulations.

[19] Note that for expositional reasons, Table B2, columns 1 and 2, are identical to Table B1, columns 3 and 6, respectively.

treated clusters, the rejection rates goes up to about 12 percent for 5 treated clusters, 29–34 percent for 2 treated clusters, and 72–79 percent for 1 treated cluster. As opposed to the individual level data, the difference-in-differences is mildly preferred to the triple difference estimator (in both its forms), and there is marginal improvement by aggregating, but typically only by a few percent, for both estimators, which unfortunately is not very helpful considering the overall scale of over-rejection. Further, there is virtually no difference between the triple difference as difference-in-differences and the full triple difference estimator. This is of course partly true because they use the same cluster asymptotics. Had we used robust (for instance White HC1), it is likely that the triple difference as difference-in-differences would look more like the difference-in-differences, as it would be based on regular asymptotics and they have the same degrees of freedom.

When we introduce true, simulated effects, the same pattern as individual level data arises. The triple difference, in both its forms, has higher power. With 25 treated clusters the rejection rates are 31 percent to 10 percent for a 2 percent effect, and 95 percent to 40 percent for a 5 percent effect. However, as the number of treated clusters decrease to 5, the difference also decreases, with 21 percent to 16 percent for a 2 percent effect, and 63 percent to 32 percent for a 5 percent effect. The difference becomes even smaller for 2 and 1 treated clusters.

Overall, aggregation has only minor consequences for false positives, but major consequences for power. This is seen by comparing Tables B2 and B3. Even in the case where the asymptotics seems reasonable, i.e., 25 treated clusters, the consequences with a 2 percent effect is a reduction in rejection rates from 21 percent to 10 percent for the difference-in-differences, and 55 percent to 31 percent for the triple difference. For a 5 percent effect the reduction is from 69 percent to 40 percent for the difference-in-differences, and 99.8 percent to 95 percent for the triple difference.

To conclude, for individual level data, clustered errors, and five or more treated clusters, the triple difference typically performs slightly better than the difference-in-differences, with the reverse being true for 1–2 treated clusters, when it comes to false positives. However, the differences are marginal compared to the overall issue of over-rejection. When it comes to power, the triple difference outperforms the difference-in-differences, often by a lot, in almost all cases. Aggregation does not solve much, and comes at a large cost in terms of power. It is also noteworthy that there is close to no difference between the full triple difference and the triple difference as difference-in-differences, either in terms of false positives or power. Researchers considering the triple difference should rest assured that in optimal cases with many (treated) clusters, the triple difference is typically at least as good as the difference-in-differences, and often much better. But beware that it suffers almost equally to the difference-in-differences estimator in the presence of few treated clusters and serially correlated errors.