Brown, Annette N.; Wood, Benjamin Douglas Kuflick

Article

Which tests not witch hunts: A diagnostic approach for conducting replication research

Economics: The Open-Access, Open-Assessment E-Journal

Provided in Cooperation with:
Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

Suggested Citation: Brown, Annette N.; Wood, Benjamin Douglas Kuflick (2018) : Which tests not witch hunts: A diagnostic approach for conducting replication research, Economics: The Open-Access, Open-Assessment E-Journal, ISSN 1864-6042, Kiel Institute for the World Economy (IfW), Kiel, Vol. 12, Iss. 2018-53, pp. 1-26, https://doi.org/10.5018/economics-ejournal.ja.2018-53

This Version is available at:
https://hdl.handle.net/10419/181452

Economics
The Open-Access, Open-Assessment E-Journal

Vol. 12, 2018-53 | August 16, 2018 | http://dx.doi.org/10.5018/economics-ejournal.ja.2018-53

Which tests not witch hunts: a diagnostic approach for conducting replication research

Annette N. Brown and Benjamin Douglas Kuflick Wood

Abstract

Replication research can be used to explore original study results that researchers consider questionable, but it should also be a tool for reinforcing the credibility of results that are important to policies and programs. The challenge is to design a replication plan open to both supporting the original findings and uncovering potential problems. The purpose of this paper is to provide replication researchers with an objective list of checks or tests to consider when planning a replication study. The authors present tips for diagnostic replication exercises in four groups: validity of assumptions, data transformations, estimation methods, and heterogeneous impacts. For each group, the authors present an introduction to the issues, a list of replication tests and checks, some examples of how these checks are employed in replication studies of development impact evaluations, and a set of resources that provide statistical and econometric details. The authors also provide a list of don'ts for how to conduct and report replication research.

(Published in Special Issue The practice of replication)

JEL C10 B41 A20

Keywords Replication; validation; data transformation; estimation methods; reanalysis; theory of change; assumptions

Authors

Annette N. Brown, FHI 360, Washington, DC, USA, abrown@fhi360.org, ABrown@fhi360.org

Benjamin Douglas Kuflick Wood, International Initiative for Impact Evaluation, Washington Office, USA

Citation Annette N. Brown and Benjamin Douglas Kuflick Wood (2018). Which tests not witch hunts: a diagnostic approach for conducting replication research. *Economics: The Open-Access, Open-Assessment E-Journal*, 12 (2018-53): 1–26. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-53

Received September 8, 2017 **Published as Economics Discussion Paper** September 28, 2017
Accepted July 14, 2018 **Published** August 16, 2018

1 Introduction

While most researchers accept the scientific premise for replication, replication studies, especially *internal* replication studies where replication researchers work with the data from the original study, often incite acrimonious responses from original authors. Gertler, Galiani, and Romero (2018) cite three such original author responses, where in two cases the replication study actually supported the original results. Gertler, Galiani, and Romero posit that it is "the current system that makes original authors and replicators antagonists" and they provide survey evidence that economics journal editors are more likely to publish replication studies that overturn results than that confirm results. The implication, also stated by one set of original authors cited by Gertler and colleagues, is that the incentives of replication researchers to publish could lead them overstate the criticism of the original article. Gertler and colleagues support this implication using the example of the "worm wars" debates[1] and reporting that "several independent scholars speculated that assumptions made by the replicators had more to do with overturning results than with any scientific justification".

Certainly one role of replication should be to explore findings that one considers questionable. However, as we argue in Brown, Cameron and Wood (2014), replication can also be a tool for reinforcing the credibility of findings that one considers important to policies and programs. The challenge is to design a replication plan open to both supporting the original findings and uncovering potential problems. When looking for tests to run, researchers instinctively look for what seems "wrong". In some ways the scientific process is grounded in having scientists serve as devil's advocates. This is not the same as striving to overturn results. Our experience is that many replication researchers are motivated to strengthen the evidence. To that end, we present here a diagnostic approach to conducting replication research to help replication researchers design plans more along the lines of a check list than a "witch hunt".[2]

We were inspired by the feedback we received from the replication researchers we worked with under the International Initiative for Impact Evaluation's (3ie's) Replication Program.[3] Under this program, researchers apply for grants to conduct replication studies of papers that are pre-selected to a "candidate studies list". The selection of studies is based on a few factors, mostly around how important the study has been or is likely to be for programs and policy. In the most recent grant round, a funder selected the studies for the list based on those it uses in determining which programs to fund. As described in Wood and Brown (2015), the program requires grantees to submit replication plans for review and then online posting. The replication plans should outline which exercises or tests beyond the pure replication would help validate the robustness and meaningfulness of the published results for informing policy. Some researchers commented to us that they did not know where to start, especially in the absence of the data and supporting documentation from the original study.

---

[1] See Section 7 of this article for more discussion of worm wars.

[2] See Blattman (2015) and Zimmerman (2015) for examples where this term is used in the context of assuming bad incentives on the part of replication researchers.

[3] Annette Brown directed the 3ie Replication Program from its establishment through July 2016. Benjamin Wood managed the program from 2012 until Brown's departure and directed the program from then through June 2018.

Our idea to develop something like a diagnostic tool was inspired by risk-of-bias assessment tools used for systematic reviews. See Waddington et al. (2017) for an overview of such tools. We had one grantee, Fernando Martel García, who employed a risk of bias assessment as part of a replication study. The assessment was useful for evaluating the strengths and the weaknesses of the original study according to pre-determined standards. It was not so useful for identifying measurement and estimation checks, however, as many of the threats to bias identified in risk-of-bias tools, such as whether treatment status is blinded, concern things that cannot be changed once the data are collected.

In compiling the tips and examples below, we started by looking at existing risk of bias tools but then relied in large part on what we have seen in replication research generally and in the studies specific to development impact evaluation. We group the empirical tests in four sections: validity of assumptions, data transformations, estimation methods, and heterogeneous outcomes. We also include a list a list of don'ts for conducting and reporting replication research.

The diagnostic approach in the replication context

The diagnostic approach to replication does not cover all the exercises a replication researcher may want to conduct. To put these tests in the larger context of replication research, we use our Brown, Cameron and Wood (2014) taxonomy, which defines pure replication, measurement and estimation analysis, and theory of change analysis. One reason we prefer this taxonomy is because the research that motivates our work – development impact evaluation – crosses many disciplines that have different replication taxonomies. When we developed the taxonomy, we intended for the self-evident category names to be interpretable by scientists and social scientists. Also, we delineated the categories to focus on the issues to be examined (measurement, estimation, theory of change) rather than blanket conclusions to be drawn (e.g. reproducible, robust), as those labels can be a source of conflict that can inhibit replication's role in the scientific process.

We can partially map our taxonomy to some of those used by economists. What we call pure replication is called reproduction (Reed, 2017), verification (Clemens, 2017), and pure replication (Hamermesh, 2007). What we call measurement and estimation analysis is similar to robustness analysis—same data set (Reed, 2017) and reanalysis (Clemens, 2017). These categories and labels refer to the use of the original article's data, that is, same population and same sample[4]. We should note that some of th tips do require the use of other data, but these exercises are not statistical replication (Hamermesh, 2007) or external replication (conducting the same study on a different population). More generally, our use of the generic term replication is the same as described by Duvendack, Palmer-Jones, and Reed (2017), a "study whose main purpose is to determine the validity of one or more empirical results from a previously published study".

[4] We do include in measurement and estimation analysis, and in the data transformations group of diagnostic tests, consideration of outliers and excluded observations, which Clemens includes in his extensions category.

The motivation to identify diagnostic tests presumes that that replication researcher intends to go beyond pure replication. The first grouping in the diagnostic approach, validity of assumptions, includes checks on the setup of the study and the choice of estimation methods. The tests do not re-analyze the data to answer the original study's research question, but they are related to estimation analysis as they use the data to check the assumptions behind the estimation conducted in the original. Many of the tests in the data transformations group relate to measurement analysis, as the creation of key variables often involves transforming the data, and imputation of missing values introduces new measurement. For diagnostic purposes, replication researchers may also want to check how transformations that change the sample influence the results, so we have included these checks in this group. The third group, estimation methods, is a subset of estimation analysis.

What we call theory of change analysis does not have a direct match in other taxonomies, and in its full meaning extends beyond a diagnostic approach to replication. That is, replication researchers who conduct extensive theory of change analysis typically initiate their studies with a non-neutral view that the theory of change is different than proposed in the original article. We do include a section here on testing for heterogeneous outcomes, which may sometimes look like diagnostic tests (e.g. wanting to check whether the results for women are different than for men) but is more often related to theory of change analysis. Regardless of whether the tests are proposed for diagnostic or theory of change purposes, replication researchers should discuss the selection of sub-groups they test in terms of theory of change, so as not to risk "fishing". There are other replication exercises used in theory of change analysis, most notably specification tests (separate from estimation methods tests), that we do not include here as diagnostic tests. Even in a purely diagnostic approach, however, theory of change considerations can, and should, inform the selection and application of the diagnostic tests. See for example, Bärnighausen et al. (2017) on exploring the theory behind assumptions for quasi-experimental analysis and Caliendo and Kopeinig (2008) for using theory to inform the choice of variables in a matching specification.

These suggested exercises are not meant to cover all the possible approaches to conducting a replication study. Instead they are intended as a neutral checklist that can help replication researchers identify useful ways of providing additional information about an original article's results. Ideally both theory and context inform the exercises chosen by replication researchers in all four sections. For each of the four sections, we provide an overview, list some suggested tips, and then give examples from 3ie-funded replication research.[5] We do not provide detailed statistical descriptions of suggested tests. Rather we suggest some selected resources with keywords at the end of each section.

[5]In some cases where there is an accepted manuscript we cite both the accepted manuscript and the working paper, as the working papers contain more results.

2 Validity of assumptions

In conducting empirical research, the theories that we apply and the empirical methods that we employ are based on assumptions. Often academic debates concern assumptions, especially as some assumptions are a matter of opinion or perception. There are ways to explore or test many of the assumptions we use in empirical analysis. When authors do not report the results of assumption tests, it may just reflect space constraints, but sometimes there are applicable tests they do not perform. And sometimes seeing the results of assumptions tests can help us to better interpret the results of the analysis. For impact evaluations, the assumptions that receive the most attention are those required in claiming that the identification strategy achieves internal validity.

For randomized controlled trials (RCTs), identification comes from random assignment, and we typically test whether the random assignment has produced a valid comparison group by looking at group equivalence for observable variables. It is important to examine group equivalence for RCTs, as even carefully designed studies can suffer from randomization failure (King, Nielson, Coberley, Pope, and Wells, 2011) or selective attrition. Anderson (2013) reanalyzes the data from the Malesky, Schuler, and Tran (2012) randomized experiment and shows that pre-treatment differences between the treatment and control groups can explain the estimated treatment effect. Some argue that balance tests are not appropriate for RCTs if there were no recognized threats to the randomization process. See McKenzie (2017) for a useful discussion. Researchers often use statistical balance tests to assess group equivalence, but sometimes visual comparisons and other tests can also be useful, especially if a replication researcher is interested in the distributions of the characteristics and not just the means. Another consideration is whether the group equivalence assumption holds for the relevant levels of analysis or for the final analysis dataset as opposed to the initial survey or recruitment sample.

For as-if random and other statistical designs, identification requires specific assumptions depending on the empirical approach. Bärnighausen et al. (2017) provide a useful review of the assumptions required for the five main quasi-experimental approaches and then describe tests that can be performed to test those assumptions. As some of these tests are recent methodological innovations, a replication study could usefully apply one or more of these tests to a study that was not able to benefit from the tests when it was conducted. The different tests provide different kinds of information; for example, some tests can only falsify an assumption but not validate it (such as balance-type tests for the continuity assumption for regression discontinuity design). Rothstein (2017) is a recent example of a replication study that tests the assumptions of a natural experiment design, including using a placebo test, and Chetty, Friedman, and Rockoff (2017) provide a useful response. In their study of Nunn and Qian's (2014) highly cited article suggesting that food aid causes conflict in recipient countries,

Christian and Barrett (2017) use placebo tests, randomization inference tests, and Monte Carlo simulations to argue that the exclusion restriction supporting Nunn and Qian's panel instrumental variables estimation strategy is questionable.

Identification assumptions are not the only assumptions that matter, though. For example, Alevey's (2014) evaluation of the impacts of the Milliennium Challenge Corporation's roads investments in Nicaragua implicitly assumes that prices do not change between the two locations connected by better roads. He measures the benefits using the travel time and cost and

the number of travelers. Parada (2016) points out that this assumption is not likely to hold when one endpoint is an interior urban area and the other is an isolated coastal area. He tests the assumption empirically and then re-analyzes the benefits of the investment taking into account relative price changes.

See Table 1 for suggested tips for validating assumptions in a replication study.

Table 1: Tips for exercises to validate assumptions

☐ Test balance between treatment and control or comparison for relevant variables
☐ Test balance at applicable units of analysis or for analyzed subsets of the data
☐ Examine the size of differences in observable characteristics between treatment and control or comparison
☐ Use outside data to explore equivalence of groups or clusters used in the study
☐ Run placebo tests, especially for natural experiment designs
☐ Explore assumptions visually, especially distributions or trends over time
☐ Identify important untested assumptions for chosen estimation methods and test using accepted methods
☐ Run randomization inference and randomization interference tests

Examples

In their replication study of Galiani and Schargrodsky's impact evaluation of property rights for the poor, Cameron, Whitney, and Winters (2015 and 2018) recognize that the balance tests reported in the original article are applied to the full sample of data (1082 observations) while the main analysis of the original study is conducted on a subsample of 300 observations. In the replication study, Cameron et al. (2015) the results of balance tests on the same 300 observations used in the main analysis, focusing on the pre-treatment characteristics of the parcels of land, which may be considered to have direct bearing on the outcomes of interest. They find statistically significant differences in three of the four parcel characteristics, whereas only one of the four is different on average for the full sample. Nonetheless, Cameron et al. find that these pre-treatment differences in the main analysis sample do not change the main findings of the original paper.

Donato and Garcia Mosqueira (2018) study Björkman and Svensson's (2009) field experiment testing the effect of community-based monitoring on health and education outcomes. They observe in the original study that the authors report pretreatment balance on a number of factors at the facility and community level but do not report any information about balance in household characteristics. The theory of change for the intervention is that it motivates households to monitor service providers, so Donato and Garcia Mosqueira look at the pretreatment balance of household characteristics that may have an impact on the relevant outcomes. They find that the treatment and control households are balanced. Donato and Garcia Mosqueira also realize that the dataset includes pretreatment observations for the immunization indicators Björkman and Svensson use at endline to estimate a positive effect of the intervention. Donato and Garcia Mosqueira test pretreatment balance between the treatment and control groups and find that prior to the intervention, the treatment group had statistically significantly better outcomes for three of five indicators. After further examination of the trends

in immunization rates, they conduct difference-in-difference analysis, which suggests that the program had no measurable impact on immunization rates.

The Bowser (2015) replication study of the Dercon et al. (2009) impact evaluation of roads in Ethiopia provides a good example of testing an assumption not directly related to the identification strategy. The growth model used for estimation by Dercon et al. assumes that access to technology, capital stock accumulation, and consumption levels change very slowly over time, such that the observed initial period values approximately equal the values in the prior period. Bowser tests these assumptions by using a multivariate test of mean equality of each variable across rounds. He reports that in all cases, the null hypothesis of equality is rejected, which suggests that the assumption is invalid. Bowser thus re-analyzes the data employing an estimation technique that does not rely on the assumption and finds that some results from the original study are strengthened while others are weakened. Table 2 provides a selection list of resources for validating assumptions.

Table 2: Resources for validating assumptions

Citation	Key words
Bärnighausen et al. (2017) "Quasi-experimental study designs series – Paper 7: assessing the assumptions"	Quasi-experimental designs, tests for weak instruments, overidentification tests, exclusion restriction, instrument validity, monotonicity assumption, balance tests, manipulation tests, pre-treatment trends, treatment reversals,
Bruhn and McKenzie (2009) "In pursuit of balance: randomization in practice in development field experiments"	Balance tests
King et al. (2011) "Avoiding randomization failure in program evaluation, with application to the Medicare Health Support Program"	Control of variability, levels of randomization, size of treatment arms, design errors
De la Cuesta and Imai (2016) "Misunderstandings about the regression discontinuity design in the study of close elections"	As-if-random assumption, continuity, extrapolation, multiple testing, placebo test, sorting
Roodman (2009) "A note on the theme of too many instruments"	Generalized method of moments estimators, Hanson test of instruments' joint validity, overfitting
Rothstein (2010) "Teacher quality in educational production: tracking, decay, and student achievement"	Placebo test
Helland and Tabarrok (2004) "Using placebo laws to test 'more guns, less crime"	Placebo test
Imbens (2015) "Matching methods in practice: three examples"	Plausibility of unconfoundedness
Calonico et al. (2015) "Optimal data-driven regression discontinuity plots"	Detection of discontinuities
Aronow (2012) "A general method for detecting interference between units in randomized experiments"	Randomization inference, randomization interference

3 Data transformations

Researchers often transform their data to prepare it for analysis. These transformations all involve choices. Data transformations include actions such as deleting or weighting outliers, imputing missing values or dropping observations that have missing values, and using data to construct new variables. While these choices are inevitable, they are not always documented by researchers and rarely reviewed by referees. Replication research can usefully explore the robustness of study results to the choices made in transforming the data for estimation.

There are several approaches to handling missing data. Researchers may choose to drop observations with missing values, assign all missing values the same value (based on an assumption, for example, about why a response was not given), impute missing values using variable means, or use other imputation methods. Lall (2016) replicates a large number of empirical political science studies using multiple imputation instead of listwise deletion for missing values and finds that this changes the results for almost half of the studies. Unless there is a clear explanation for missingness that points to an assigned value or method, replication can test the robustness of the original results to alternative missing data techniques.

It may also be useful to look at excluded observations. Researchers regularly identify outliers, based for example on statistical tests, distributional analysis, or contextual knowledge. After identifying these outliers, researchers make choices about whether and how to transform them. They may delete them, winsorize them, or use other tools to transform them. Replication analysis can reconsider the assumptions implicit in the transformation and can test the robustness of the results to these transformations.

A third area for analysis of data transformation is variable construction. To create variables to represent the concepts being studied, researchers often construct new variables using values from other variables. Depending on the concept being measured and the planned estimation strategy, researchers may sum values, construct indexes, convert categorical variables to binary variables, weight values across observations, and so on. These choices always involve some element of subjectivity. Even for the most straightforward construction of a truly quantitative variable such as income, for example, researchers must decide how to treat in-kind transactions. Replication researchers can reconsider the theories and assumptions supporting data transformation decisions and test the robustness of results to the constructions used. Replication researchers can also use alternate data to test the consistency of the measurement for the same observations or to test the robustness of the results to values measured using alternate data. For example, in their replication study of the effect of corruption on election results, Goel and Mazhar (2015) argue that corruption is a difficult concept to measure and use a corruption index from another, better justified, data source to test the robustness of the results from the original study. Scherer (2015) replicates the OECD fragility index and finds that more than half of the countries measured by the OECD are misclassified.

Table 3 provides suggested tips for assessing data transformations as part of a replication study.

Table 3: Tips for data transformation exercises

☐ Employ alternative imputation methods for missing values to test robustness
☐ Use an alternative outlier drop rule to test robustness
☐ Explore the impact on the results of any dropped observations
☐ Decompose constructed variables to understand the implications of the composition and weights
☐ Consider different constructions supported by theory or qualitative analysis
☐ Use alternate data for key variables to test robustness

Examples

In the Basurto et al. (2015) replication study, the researchers note that the development of the HHH2009 dataset used by Cattaneo et al. (2009) included two different approaches to handle missing values. The first approach was to impute values using Dummy Adjustment Imputation, which is declared in the supporting documentation. Basurto et al. also find that for three constructed variables in HHH2009 – per capita cash transfers from government programs, total per capita value of household assets, and total per capita consumption – the Arithmetic Mean Imputation method was used to fill in missing values for the original variables used in the construction. Basurto et al. test the robustness of the published result by using the Multiple Imputation method for missing values. They try three specifications for their multiple imputation calculations and find very similar results across all three. Ultimately, the researchers demonstrate that the original result – adding concrete floors to households improved children's health – is robust to different approaches to imputing missing data. Basurto et al. include an appendix reviewing the literature around imputation methods.

Djimeu, Korte and Calvo (2015) in their replication study of Bailey et al. (2007) look at whether the missing data due to loss to follow up could have changed the findings from the study. The original study estimates the effect of male circumcision on HIV incidence. The replication researchers estimate what the HIV outcomes would need to be among those lost to follow up if those observations added to the study data would cause the study findings to be significantly changed. They conclude that the difference between the lost to follow up group and the study group would have to be implausibly high for the study findings to be changed. Djimeu, et al. conclude that the original results are not sensitive to missing data.

Kuecken and Valfort's (2018) replication study examines several elements of the Reinikka and Svensson (2005) paper, including the exclusion of certain schools from the analysis dataset. The original study demonstrates how an anti-corruption newspaper campaign focused on schools increased student enrollment and learning. Kuecken and Valfort question the decision by the authors to exclude from their analysis a limited number of schools that recorded a decrease in student enrollment. They note that the footnote in the original study explains that these schools experienced reductions in enrollment due to "idiosyncratic shocks", and thus the original authors argue that such shocks should not be systematically correlated with the explanatory variable. After reintegrating the dropped schools into the sample, however, Kuecken and Valfort find the published statistical significance of the change in enrollment is sensitive to the exclusion of these schools.

The Iversen and Palmer-Jones (2018) replication study of Jensen and Oster's (2009) impact evaluation of the effect of the introduction of cable TV on women's autonomy in India includes a detailed examination of the construction of the index variables that the original authors use to measure their primary outcomes. These indexes aggregate information from multiple survey questions, each designed to measure something different about individual or household situations. Iversen and Palmer-Jones draw on theoretical work on female empowerment to analyze the interpretation and inclusion of each of these questions in a composite index. When they construct alternate variables based on the theory, they find that the statistical significance of several results changes. Table 4 presents selected resources for exploring data transformations.

Table 4: Resources for data transformation exercises

Citation	Key words
Alsop et al. (2006) "Empowerment in practice: analysis to implementation"	Analytic framework, methodological issues in measuring empowerment
Ceasar de Andrade et al. (2013) "Evaluation of the reliability and validity of the Brazilian Healthy Eating Index Revised"	Index validation, content validity, construct validity
Kilic et al. (2017) "Missing(ness) in action: selectivity bias in GPS-based land area measurements"	Missing geographic observations
Lall (2016) "How multiple imputation makes a difference"	Data imputation techniques, multiple imputation
Matern et al. (2009) "Testing the Validity of the Ontario Deprivation Index"	Index validation, poverty measurement
Samii (2016) "Inverse covariance weighting versus factor analysis"	Index construction, inverse covariance weighting, factor analysis

4 Estimation methods

Statisticians across all fields of study are innovative and prolific, and the result is that many different methods have been developed to do some of the same things. Methods within disciplines also evolve over time. Original studies often report robustness tests or sensitivity analysis to various estimation techniques, but replication studies can fill in where this analysis is missing or build on the analysis using newer methods. Moundigbaye, Rea, and Reed (2018) find, for example, that "while OLS with cluster robust standard errors is widely used by applied researchers, our experiments find that it performs relatively poorly on both efficiency and inference grounds for the small to moderately-sized panel datasets" they studied (p. 28). In this section, we are not talking about testing alternate specifications, which might be part of a theory of change analysis. We are talking examining alternate estimation methods for testing the same relationships as in the original study.

One way to examine the robustness of a study's results to different estimation methods is to employ the methods of another discipline. For example, for epidemiological research, a replication study might apply econometric approaches or vice versa. We see these two disciplines overlap more often as epidemiologists do more implementation science research to understand whether and how health programs work, and economists (and other social scientists using econometric methods) are conducting their own RCTs of health-related interventions. Powell-Jackson et al. (2018) discuss the differences between the two disciplines and suggest that each can learn from the other. The replication studies that set off the "worm wars" (Aiken et al., 2015 and Davey et al., 2015) are examples of epidemiologists using their methods to reanalyze an economics study of a health intervention. We show an example of an econometric replication study of an epidemiology paper below.

Quasi-experimental methods typically require researchers to make more decisions about how to employ their estimation methods. One example is when matching is used as the identification strategy. There are multiple matching methods that can be used, such as exact, coarsened exact, and propensity score, and there are choices to make within methods, such as which variables to include in the propensity score regression. Imbens (2015) argues about matching estimators that "the results should not be sensitive to the choice of reasonable estimators" (p. 374). Smith and Todd (2005) apply multiple matching techniques to the National Supported Work data famously analyzed by LaLonde (1986) and find that the results are quite sensitive to the estimator chosen. Lampach and Morawetz (2016) follow the steps outlined by Caliendo and Kopeinig (2008) to reanalyze the data of Jena et al. (2012), who use propensity score matching to test the effects of Fair Trade coffee certification. Lampach and Morawetz compare the results from nearest neighbor matching, used by Jena et al., to Mahalanobis-distance matching and genetic matching and report that the other models do not find the significance influence of certification on consumption found in by Jena, et al.

Another example of different approaches to a quasi-experimental estimation method is regression discontinuity design. Button (2017), noting that "regression discontinuity design literature has improved significantly", conducts a replication study of Lee, Moretti, and Butler (2004) using more advanced regression discontinuity techniques and finds that the original results are robust to the newer techniques.

There are also some checks on estimation methods that are closer to being corrections. A standard example is correcting for clustered standard errors if the original study uses a clustered design but did not calculate corrected standard errors. Another example is adjusting for multiple hypothesis testing, about which Lakens (2016) provides a useful discussion.

Table 5 provides some suggested tips for checking estimation methods.

Table 5: Tips for checking estimation methods

☐ Run additional robustness tests for key parameter or specification choices in the estimation strategy
☐ Explore estimation strategies from other disciplines with applicable approaches, especially in cases whether the other disciplines sometimes analyze similar questions
☐ Apply newly available techniques for an estimation strategy
☐ Check for the correct application of estimation strategies given the set-up of the study

Examples

In the Korte, Djimeu and Calvo (2018) replication study of Bailey et al.'s 2007 RCT of male circumcision in Kenya, the replication researchers use econometric methods to test the same relationships that the original paper explores with epidemiology methods. (See also Djimeu et al., 2015.) Korte and team run ordinary least squares, fixed effects, and instrumental variable regressions to estimate the effect of male circumcision on HIV incidence. The fixed effects model helps to control for unobserved individual heterogeneity by exploiting the panel nature of the data. The instrumental variable estimation uses the random assignment as the instrument and includes other possible explanatory variables for the decision to circumcise. These alternate estimation methods produce very similar estimates as reported by Bailey et al. and confirmed by Korte et al.'s pure replication, but uncover evidence suggestive of risk compensation, a policy relevant concern ruled out by Bailey et al.

Cameron et al. (2015 and 2018) explore Galiani and Schargrodsky's (2010) assessment of the effects of land titling on urban poverty. The original authors focus their study on a number of outcomes of interest, including: housing investment, household structure, human capital accumulation, access to credit and labor earnings. As one of the replication checks, the replication researchers determine the sensitivity of the results when accounting for multiple hypothesis testing. Cameron et al. find the statistical significance originally reported for the disaggregated household investment variables is generally robust to correction for multiple hypothesis testing.

Carvalho and Rokicki (2018) explore the robustness of the results from the exact matching strategy employed by Lim et al.'s 2010 impact evaluation of the India Janani Suraksha Yojana (JSY) conditional cash transfer programme. Carvalho and Rokicki estimate three different propensity score matching models, first using the same covariates in the propensity score regression as in the exact matching algorithm, then adding additional covariates, and then adding district fixed effects. In all cases, the results are very similar to those in the original. Cameron et al. (2015 and 2018) also look at alternative matching methods in their replication study. The original study uses manual matching, and Cameron and team employ propensity score matching as a robustness check. In their propensity score analysis, they add two covariates that are also arguably time invariant, gender and education of the original squatter. The results from these alternative estimation strategies are consistent with those in the original article.

Table 6 suggests some resources on estimation methods.

Table 6: Resources for checking estimation methods

Citation	Key words
Stuart (2010) "Matching methods for causal inference: a review and a look forward"	Epidemiology, distribution of covariates, closeness, distance, nearest neighbor, weighting, common support, diagnosing matches
Imbens and Wooldridge (2009) "Recent developments in the econometrics of program evaluation"	Econometrics, estimation inference, Ruben causal model, average treatment effects, randomized experiments, regression models, propensity score
Imbens and Kalyanaraman (2012) "Optimal bandwidth choice for the regression discontinuity estimator"	Regression discontinuity, local linear regression, optimal bandwidth selection, cross validation, simulation study
Anderson (2008) "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training projects"	Multiple hypothesis testing, false discovery rate, familywise error rate, multiple comparisons, summary index
Kling et al. (2007) "Experimental analysis of neighborhood effects"	Multiple hypothesis testing, summary indices
Ozler (2014) "Obesity may not have dropped among children, but it almost certainly increased among the elderly"	Multiple hypothesis testing techniques, Bonferroni correction, family-wise error rate, free step-down resampling
Romano and Wolf (2005) "Stepwise multiple testing as formalized data snooping"	Bootstrap, data snooping, familywise error, multiple testing, stepwise method.
Young (2017) "Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results"	Multiple hypothesis testing, bootstrap, randomization tests, joint tests, omnibus randomization test
Bowers and Coopers (n.d.) "10 things to know about cluster randomization"	Information reduction, cluster sizes, within-cluster spillovers, power analysis
Caliendo and Kopeinig (2008) "Some practical guidance for the implementation of propensity score matching"	Matching algorithm, matching quality assessment, sensitivity analysis
Imbens (2015) "Matching methods in practice: three examples"	Sub-classification on the propensity score, trimming, pre-processing methods
Calonico et al. (2014) "Robust nonparametric confidence intervals for regression-discontinuity designs"	Bandwidth selectors, bias-corrected estimator
Moundigbaye et al. (2018) "Which panel data estimator should I use? A corrigendum and extension"	Monte Carlo simulations, bootstrapping, data generating process, panel-corrected standard errors, feasible generalized least squares

5 Heterogeneous outcomes

The replication study by Cervellati et al. (2014) of Acemoglu et al.'s (2008) study "Income and Democracy" provides a simple explanation of why it is important to conduct subgroup analysis. In a linear estimation framework, the question is whether different subgroups have not just different intercepts, which might be controlled for using dummy variables or fixed effects, but

also different slopes, indicating a different kind of effect, or a different theory of change. In that study, the replication researchers conduct subgroup analysis and find that there indeed are effects of income on democracy, contrary to the findings from the analysis of the full sample of countries, but that those effects are different based on whether a country has ever been colonized. Cervellati et al. use theory to select the subgroups they consider, and in fact, draw on the discussion in the original article for this analysis.

For policy making, it is often critical to understand the heterogeneous impacts of interventions or programs. Imai and Ratkovic (2013) argue that estimating treatment effect heterogeneity is important for "(1) selecting the most effective treatment from a large number of available treatments, (2) ascertaining subpopulations for which a treatment is effective or harmful, (3) designing individualized optimal treatment regimes, (4) testing for the existence of lack of heterogeneous treatment effects, and (5) generalizing causal effect estimates obtained from an experimental sample to a target population." (p. 1)

There are different reasons why original authors might not conduct subgroup analysis or test for heterogeneous outcomes, and replication researchers should carefully consider them when embarking on subgroup analysis. One is simply sample size, particularly if the original sample is not large. Simulation methods, such as randomization inference, can be useful for hypothesis testing in these cases. Another is the desire to maintain the statistical assumptions afforded by randomized assignment, which do not apply if the random assignment did not stratify to subgroups. Replication researchers should be clear that their findings are descriptive and not causal if the randomization was not stratified. They may also want to conduct balance tests to explore the pre-treatment comparability of the treatment and control observations in the subgroup as part of the analysis. And finally, subgroup analysis introduces the multiple comparison, or multiple hypothesis tests, issue discussed in the previous section. Replication researchers should apply the appropriate corrections.

Table 7 suggests the steps for exploring heterogeneous outcomes.

Table 7: Tips for heterogeneous outcomes exercises

☐ Identify theoretically or clinically relevant subgroups and check whether heterogeneous impacts are tested for these subgroups
☐ Test for heterogenous impacts for relevant subgroups taking into account sample size, identification assumptions, and multiple hypothesis testing

Examples

Carvalho and Rokicki (2018) reexamine an evaluation of Janani Suraksha Yojana (JSY), a large-scale conditional cash transfer in India that incentivizes women to use formal birthing facilities. Lim et al. (2010) use a range of estimation techniques, including exact matching and difference and difference analysis, to estimate the effect of the program on uptake and health. While the original analysis examines state-level health outcomes, the authors chose to focus their coverage outcomes at the regional level. After reproducing the original results, the replication researchers extend the coverage outcomes to the state level. Their sub-group reanalysis shows a wide amount of heterogeneity in state level coverage outcomes, especially in

reproductive health coverage indicators. These findings suggest future researchers of the JSY program should account for state level heterogeneity in their evaluations.

Wood and Dong (2015) re-examine an agricultural commercialization evaluation, where the intervention included specific export oriented crops promotion, the easing of transportation constraints, and a formalization of the crop sales process (Ashraf, Giné, and Karlan, 2009). The original authors test heterogeneous impacts by splitting their sample according to farmers' previous export crop producer status, focusing specifically on the three crops promoted in the intervention. They find that the intervention benefits those who did not previously grow export crops. Based on value-chain theory, the replication researchers hypothesize that farmers' previous participation in markets, regardless of domestic or exports crops, should distinguish those who are more likely to benefit. They test for these heterogeneous impacts and find that farmers who did not previously sell at markets benefited from the intervention, while those who did sell at markets did not benefit. Wood and Dong conclude from these results that policymakers should focus on encouraging farmers to enter the value chain by selling their crops at markets rather than incentivizing the production of specific crops.

Iversen and Palmer-Jones (2018) also use theory to motivate testing for heterogeneous impacts. The original paper examines how the expansion of cable access in rural India influenced a number of women's rights (Jensen and Oster, 2009). The replication researchers explore the theory of change by examining the mechanisms leading to the change in the observed outcomes in more detail. Ultimately, the reanalysis suggests that cable TV access may influence certain women's rights more than others, especially women with some previous educational attainment. The replication researchers advise modifying the policy recommendations stemming from this research and further investigating the influence of this intervention before promoting it to policymakers.

Cameron, Whitney and Winters (2015) provide an assessment of the theory of change as described by Galiani and Schargrodsky (2010) drawing on a more extensive literature review. Based on this review, they hypothesize that the pathways for property rights to impact urban households might be different based on household characteristics. Data limitations prevent them from conducting a more extensive theory of change analysis, but they are able test for heterogeneous impacts for gender of the original squatter and education of the original squatter. They find that the treatment effects are indeed different men and women and for original squatters who completed at least primary education versus those who had not (Cameron, Whitney and Winters, 2015 and 2018). Table 8 provides a selected set of resources about heterogeneous outcomes.

Table 8: Resources for heterogenous outcomes

Citation	Key words
Khandker et al. (2010) "Handbook on impact evaluation: Quantitative methods and practices"	Linear regression framework, heterogeneous program impacts, quantile regression
Evidence in Governance and Politics (n.d.) "10 things to know about heterogeneous treatment effects"	Testing for heterogeneity, conditional average treatment effects, interaction effects
Imai and Strauss (2011) "Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the Get-Out-the-Vote campaign"	Heterogeneous treatment effects, two-step framework, post hoc subgroup analysis
Imai and Ratkovic (2013) "Estimating treatment effect heterogeneity in randomized program evaluation"	Variable selection problem, support vector machine, sampling weights
Varadhan and Seeger (2013) "Chapter 3: Estimation and reporting of heterogeneity of treatment effects"	Heterogeneity of treatment effect, clinically relevant subgroups, observational comparative effectiveness research
Cummins (2017) "Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment"	Outcome distribution, targeting of behavioural responses, rank similarity, generalizability
Athey and Imbens (2017) "The econometrics of randomized experiments"	General treatment effect heterogeneity, covariates, valid confidence intervals
Green (n.d.) "10 things to know about randomization inference"	Calculating p-values for hypothesis tests, simulations, multiple comparisons

6 A checklist of don'ts

Don't confuse general critiques of the original research with replication tests. Replication tests as covered here are re-analyses of the original data to address the original research question. Sometimes replication researchers include critiques of the original study that are not related to the replication tests they carry out with the data. Examples of such critiques are arguing that the original design of an experiment was flawed, disagreeing with the choice of, or interpretation of, studies included in the literature review, or offering an alternate explanation for the original article's results. To the extent that these critiques directly motivate the choice of replication tests to run, they are part of a replication study. But if they are not informing the replication tests, they should be set apart and clearly marked as general critiques and not replication findings.

Don't label a difference between the pure replication results and the original results an "error" or "mistake" without determining the source of the error. We present a longer discussion of this in Brown and Wood (2014). Just because the second estimate is different from the first does not make the second one necessarily right and the first one necessarily wrong. In the same vein, don't sweat the small differences. If the purpose of replication is to validate results for the purpose of using evidence to inform policies and programs, small differences in point estimates or t-statistics, even if there are many, are not likely to be crucial. Focus on identifying the

meaningful differences and then look for an explanation for why the replication results are different. Finding the explanation may involve "reverse engineering", that is deconstructing the original study step by step to figure out where the deviations occur. See Anderson (2017) for a paradigm for this process. Wood and his co-author Dong undertook such a process in their replication study of Ashraf, Giné, and Karlan (2009) to identify the source of differences in the descriptive statistics and other tables.

Don't conduct measurement or estimation analysis before conducting a pure replication, and present the results of the pure replication before presenting the other replication study results. Some people consider the pure replication to be the only "replication" test, and to the extent that the replication researcher makes a statement that the original study is "replicable" or not, it is best to limit such a statement to the pure replication results. Even if the pure replication finds no differences and thus the pure replication tables are not presented, it is important to make a clear statement about the results of the pure replication. We also recommend not using language like "success" or "fail" as the comparisons are typically more nuanced, and those words are triggers.

Don't conduct a pure replication before conducting a push button replication, assuming the original authors' code is available. A push button replication is a test of whether the original authors' code can be run on their data to produce the published results. Among other benefits, it can prevent the need later on to reverse engineer a pure replication by identifying up front issues with coding. Also do state the findings of the push button replication in the replication study write-up as some consider push button to be the only relevant "replication" test (see Wood, Müller, and Brown, 2018).

Don't include corrections in the pure replication tests. The task of a pure replication is to see if one can use the exact methods of the original study on the original dataset to produce the published results. Sometimes in this process replication researchers identify "mistakes" and then correct them, presenting the results as the pure replication results. While things like correcting standard errors for clustering or taking the log of a variable might seem completely straightforward, these changes are still part of measurement and estimation analysis. The fuzzy line is when the pure replication reveals that what was done is different than what was stated. For example, the original study states that the standard errors were corrected but the pure replication reveals that they were not. These things happen. The replication researcher can explain that the published results can be replicated in one way and then provide the results the other way. If the differences are meaningful, they should be discussed further.

Don't present extension results as replication results. Clemens (2017) makes this argument as well. Replication studies may include extensions that help clarify or refine the interpretation of findings, particularly for the purpose of policy making and program design. It is fine to include extensions in replication studies but important to distinguish them as such, separate from the kinds of replication tests listed in this diagnostic tool.

Don't forget to look for publicly available working papers for the original article being studied. Unfortunately, original authors do not always cite or include a reference to the working paper versions of their articles, but sometimes these versions include the analysis being considered by the replication researcher.

Don't present, post, or publish a replication study of any kind without first sharing it with the original authors. This is a basic professional courtesy. Sharing with them in advance does not mean that you are beholden to their comments or objections, but we recommend you give

them a few weeks to respond before you go public. We have seen cases where the replication researchers made mistakes that the original authors catch. Document in the public version of the study the date that you shared the replication study with the original authors. This can be useful if the original authors try to "scoop" corrections from the replication study.

7 Discussion and conclusion

One might ask whether these recommendations would have helped prevent the "worm wars" conflicts from the Aiken et al. (2015) and Davey et al. (2015) replication studies of Miguel and Kremer (2004). The short answer is no. The influence of the Miguel and Kremer single paper was so immense that any critiques of it were destined to produce heated debates. And debates around the original study and the replication studies also brought into play important differences in how different disciplines approach this kind of research, and those conflicts between the disciplines heightened the debates.

Perhaps the one recommendation that would have helped is the first of the don'ts in section 6. That is, there are two sets of concerns about the policy influence of Miguel and Kremer (2004): one based on criticisms both of how the original experiment was constructed and of how the original results interpreted and the other from the empirical results of replication tests. While the former concerns are certainly important to the application of the results to policy making and program design, they do not speak to whether published findings in Miguel and Kremer hold up empirically. They speak to how we should understand them. These issues are difficult to separate. Even Humphreys (2015) who provides a non-partisan, detailed review and re-replication of the original and replication studies, switches back and forth between issues of study design and interpretation and issues of empirical methods and estimates.

Researchers embark on replication study for different reasons. Sometimes it is simply a learning exercise, and often those studies do not go beyond pure replication. When they do, this diagnostic approach can be used to develop a replication plan. Many replication researchers see replication as an important part of the scientific process and conduct replication studies with the objective to independently confirm the original results or to make sure the original article included all the relevant information. In these cases, the diagnostic approach can be used to develop the replication plan. Sometimes replication researchers select an original article to study because they believe the theory of change to be different that presented or question the application of methods. While in those cases the replication plan will be designed to explore the replication researchers' concerns, the diagnostic approach may still be helpful for developing the plan or may be applied as a neutral component of the larger study.

No matter the reason for conducting a replication study, it is just as important to make supportive replication studies publicly available as it to make unsupportive replication studies publicly available. It can be challenging to publish supportive replication studies in journals, but there are some options. For economics replication studies, this journal (*Economics: The Open Access, Open-Assessment E-Journal*) is one. Another is the *International Journal for Re-Views in Empirical Economics*. The Replication Network (https://replicationnetwork.com/publishing/) has a list of economics journals that state they accept replication studies, and perhaps some of

these publish supportive replication studies. The Replication in Economics wiki (http://replication.uni-goettingen.de/wiki/index.php/Main_Page) posts information about journals that are issuing calls for replication studies and also has a database of replication studies that have been published in journals. The political science journal *Research & Politics* publishes replication studies as research notes.

There are also non-journal outlets that can be used for making supportive (and unsupportive) replication studies publicly available. The Harvard Political Science Replication Initiative is an online database of replication studies that accepts submissions from anyone as long as they follow the submission requirements. Authors can self-archive their studies on the Open Science Framework, the Munich Personal RePEc Archive, or in other archives. Many authors also have access to working paper series through their institutions or other affiliations. Even if the study is final, a working paper series can be a good way to make it publicly available. It also cannot hurt to contact the original study's journal editors to enquire whether they would consider publishing a comment. The original journal may actually be more inclined to publish a supportive comment than an unsupportive one. In all cases, however, it is important to share the replication study with the original authors before making it public.

References

Acemoglu, D., Johnson, S., Robinson, J. A. and Yared, P. (2008). Income and democracy. *American Economic Review*, 98(3): 808–842. https://www.jstor.org/stable/29730096?seq=1#page_scan_tab_contents

Aiken, A. M., Davey, C., Hargreaves, J. R. and Hayes, R. J. (2015). Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *International Journal of Epidemiology*, 44(5): 1572–1580. https://doi.org/10.1093/ije/dyv127

Alevy, J.E. (2014). Impacts of the MCC transportation project in Nicaragua. Washington, DC: Millennium Challenge Corporation.

Alsop, R., Bertelsen, M. and Holland, J. (2006). *Empowerment in practice: from analysis to implementation. Directions in development*. Washington, DC: World Bank. http://hdl.handle.net/10986/6980

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, Perry Preschool, and early training projects. *Journal of the American Statistical Association*, 103(484): 1481–1495. https://www.jstor.org/stable/27640197?seq=1#page_scan_tab_contents

Anderson, J. H. (2013). Sunshine works: Comment on the "The adverse effect of sunshine: A field experiment on legislative transparency in an authoritarian assembly". World Bank Policy Research. Working paper No. 6602. https://ssrn.com/abstract=2326778

Anderson, R. G. (2017). Should you choose to do so… A replication paradigm. Economics Discussion Papers, No 2017–79, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2017-79

Aronow, P. M. (2012). A general method for detecting inference between units in randomized experiments. *Sociological Methods and Research*, 41(1): 3–16. https://doi.org/10.1177%2F0049124112437535

Ashraf, N., Giné, X. and Karlan, D. (2009). Finding missing markets (and a disturbing epilogue): Evidence from an export crop adoption and marketing intervention in Kenya. *American Journal of Agricultural Economics*, 91(4): 973–990. https://www.jstor.org/stable/20616255?seq=1#page_scan_tab_contents

Athey, S. and Imbens, G. W. (2017). Chapter 3 – The econometrics of randomized experiments. 1: 73–140. In Duflo, E. and Banerjee, A. (Eds.) *Handbook of Economic Field Experiments*. North Holland. https://doi.org/10.1016/bs.hefe.2016.10.003

Bailey, R. C., Moses, S., Parker, C. B., Agot, K., Maclean, I., Krieger, J. N., Williams, C. F. M., Campbell, R. T. and Ndinya-Achola, J. O. (2007). Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The Lancet*, 369(9562): 643–56. https://www.ncbi.nlm.nih.gov/pubmed/17321310

Bärnighausen, T., Oldenburg, C., Tugwell, P., Bommer, C., Ebert, C., Barreto, M., Djimeu, E., Haber, N., Waddington, H., Rockers, P., Sianesi, B., Bor, J., Fink, G., Valentine, J., Tanner, J., Stanley, T., Sierra, E., Tchetgen Tchetgen, E., Atun, R. and Vollmer, S. (2017). Quasi-experimental study designs series – Paper 7: assessing the assumptions. *Journal of Clinical Epidemiology*, 89: 53–66. https://www.ncbi.nlm.nih.gov/pubmed/28365306

Basurto, M. P., Burga, R., Flor Toro, J. L. and Huaroto, C. (2015). Walking on solid ground: a replication study on Piso Firme's impact. 3ie Replication Paper 7. International Initiative for Impact Evaluation (3ie). http://www.3ieimpact.org/media/filer_public/2015/09/16/rps_7_-study_on_piso_firmes_impact.pdf

Björkman, M. and Svensson, J. (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda. *The Quarterly Journal of Economics*, 124(2): 735–769. https://doi.org/10.1162/qjec.2009.124.2.735

Blattman, C. (2015). Dear journalists and policymakers: What you need to know about the worm wars. Available at: https://chrisblattman.com/2015/07/23/dear-journalists-and-policymakers-what-you-need-to-know-about-the-worm-wars/

Bowers, J. and Cooper, J. J. (2015). 10 things to know about cluster randomization. Evidence in Governance and Politics Methods Guides. Available at: http://www.egap.org/methods-guides/10-things-you-need-know-about-cluster-randomization

Bowser, W. H. (2015). The long and short of returns to public investments in fifteen Ethiopian villages. 3ie Replication Paper 4. International Initiative for Impact Evaluation (3ie). http://www.3ieimpact.org/media/filer_public/2015/02/06/bowser-rps4-ethiopia-public-investments.pdf

Brown, A. N., Cameron, D. B. and Wood, B. D. K. (2014). Quality evidence for policymaking: I'll believe it when I see the replication. *Journal of Development Effectiveness*, 6(3): 215–235. https://doi.org/10.1080/19439342.2014.944555

Brown, A. N. and Wood, B. D. K. (2014). When is an error not an error? Development Impact. Available at: http://blogs.worldbank.org/impactevaluations/when-error-not-error-guest-post-annette-n-brown-and-benjamin-d-k-wood

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: randomization in practice in development field experiments, *American Economic Journal: Applied Economics*, 1(4): 200–232. https://www.jstor.org/stable/25760187?seq=1#page_scan_tab_contents

Button, P. (2017). A replication of 'Do voters affect or elect policies? Evidence from the US house' (Quarterly Journal of Economics, 2004). *Public Finance Review*, 46(5): 886–893. https://doi.org/10.1177%2F1091142117721739

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1): 31–72. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6419.2007.00527.x

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6): 2295–2326. https://doi.org/10.3982/ECTA11757

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, 110(512): 1753–1769. https://doi.org/10.1080/01621459.2015.1017578

Cameron, D., Whitney, E. and Winters, P. (2018). Heterogeneous effects of urban land titling: a replication of "property rights for the poor". Manuscript accepted for publication, *Journal of Development Studies.*

Cameron, D., Whitney, E. and Winters, P. (2015). The effects of land titling on the urban poor: a replication of property rights. 3ie Replication Paper 9. International Initiative for Impact Evaluation (3ie). http://www.3ieimpact.org/media/filer_public/2015/11/05/rps9-effects-of-land-titling-on-the-urban_poor.pdf

Carvalho, N. and Rokicki, S. (2018). The impact of India's Janani Suraksha Yojana conditional cash transfer programme: A replication study. Manuscript accepted for publication, *Journal of Development Studies.*

Cattaneo, M. D., Galiani, S., Gertler, P. J., Martinez, S. and Titiunik, R. (2009). Housing, health, and happiness. *American Economic Journal: Economic Policy*, 1(1): 75–105. https://www.aeaweb.org/articles?id=10.1257/pol.1.1.75

Ceasar de Andrade, S., Previdelli, Á. N., Lobo Marchioni, D. M. and Fisberg, R. M. (2013). Evaluation of the reliability and validity of the Brazilian Healthy Eating Index Revised. *Revista de Saúde Pública*, 47(4): 675–83. https://www.ncbi.nlm.nih.gov/pubmed/24346677

Cervellati, M., Jung, F., Sunde, U. and Vischer, T. (2014). Income and democracy: comment. *The American Economic Review*, 104(2): 707–719. https://www.jstor.org/stable/42920714?seq=1#page_scan_tab_contents

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2017). Measuring the impacts of teachers: reply. *American Economic Review*, 107(6): 1685–1717. https://www.aeaweb.org/articles?id=10.1257/aer.20170108

Christian, P. and Barrett, C. B. (2017). Revisiting the effect of food aid on conflict: A methodological caution. Policy Research Working Paper, No. 8171. Washington, DC: The World Bank. http://documents.worldbank.org/curated/en/723981503518830111/Revisiting-the-effect-of-food-aid-on-conflict-a-methodological-caution

Clemens, M. A. (2017). The meaning of failed replications: a review and proposal. *Journal of Economic Surveys*, 31(1): 326–342. https://doi.org/10.1111/joes.12139

Cummins, J. R. (2017). Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment. *Economics of Education Review*, 56: 40–51. https://doi.org/10.1016/j.econedurev.2016.11.006

Davey, C., Aiken, A. M., Hayes, R. J. and Hargreaves, J. R. (2015). Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology*, 44(5): 1581–1592. https://www.ncbi.nlm.nih.gov/pubmed/26203171

De la Cuesta, B. and Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, 19: 375–396. https://doi.org/10.1146/annurev-polisci-032015-010115

Dercon, S., Gilligan, D. O., Hoddinott, J. & Woldehanna, T. (2009). The impact of agricultural extension and roads on poverty and consumption growth in fifteen Ethiopian villages. *American Journal of Agricultural Economics*, 91(4): 1007–1021. https://www.jstor.org/stable/20616257?seq=1#page_scan_tab_contents

Djimeu, E.W., Korte J. E. and Calvo, F. A. (2015). Male circumcision and HIV acquisition: reinvestigating the evidence from young men in Kisumu, Kenya. 3ie Replication Paper 8. International Initiative for Impact Evaluation (3ie). http://www.3ieimpact.org/media/filer_public/2015/09/15/rps8-male-circumcision.pdf

Donato, K. and Garcia Mosqueira, A. (2018). Information improves provider behaviour: A replication study of a community-based monitoring programme in Uganda. Manuscript accepted for publication, *Journal of Development Studies.*

Duvendack, M., Palmer-Jones, R. and Reed, W.R. (2017). What is meant by "Replication" and why does it encounter resistance in economics? *American Economic Review*, 107(5): 46–51. https://www.aeaweb.org/articles?id=10.1257/aer.p20171031

Evidence in Governance and Politics. 10 things to know about heterogeneous treatment effects. Available at: http://www.egap.org/content/10-things-know-about-heterogeneous-treatment-effects

Gertler, P., Galiani, S. and Romero, M. (2018). How to make replication the norm. *Nature*, 554: 417–419. https://www.nature.com/magazine-assets/d41586-018-02108-9/d41586-018-02108-9.pdf

Galiani, S. and Schargrodsky, E. (2010). Property rights for the poor: effects of land titling. *Journal of Public Economics*, 94(9–10): 700–79. https://doi.org/10.1016/j.jpubeco.2010.06.002

Goel, R. K. and Mazhar, U. (2015). A replication of "corruption and elections: an empirical study for a cross-section of countries" (Economics and Politics 2009). *Public Finance Review*, 43(2): 143–154. https://doi.org/10.1177%2F1091142114537890

Green, D. 10 things to know about randomization inference. Evidence in Governance and Politics Methods Guides. Available at: http://www.egap.org/methods-guides/10-things-randomization-inference

Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics*, 40(3): 715–733. https://doi.org/10.1111/j.1365-2966.2007.00428.x

Helland, E. and Tabarrok, A. (2004). Using placebo laws to test "more guns, less crime". *The B.E. Journal of Economic Analysis & Policy*, 4(1): 1538–1637. https://doi.org/10.2202/1538-0637.1182

Humphreys, M. (2015). What has been learned from the deworming replications: A non-partisan view. Columbia University. http://www.columbia.edu/~mh2245/w/worms.html

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1): 443–470. https://www.jstor.org/stable/23566518?seq=1#page_scan_tab_contents

Imai, K. and Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the vote campaign. *Political Analysis*, 19(01): 1–19. https://econpapers.repec.org/article/cuppolals/v_3a19_3ay_3a2011_3ai_3a01_3ap_3a1-19_5f01.htm

Imbens, G. W. (2015). Matching methods in practice: Three examples. *The Journal of Human Resources*, 50(2): 373–419. https://ideas.repec.org/a/uwp/jhriss/v50y2015i2p373-419.html

Imbens, G. W. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3): 933–959. https://doi.org/10.1093/restud/rdr043

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1): 5–86. https://www.ifs.org.uk/publications/8669

Iversen, V. and Palmer-Jones, R. (2018). All you need is cable TV? Manuscript accepted for publication, *Journal of Development Studies.*

Jena, P. R., Chichaibelu, B. B., Stellmacher, T. and Grote, U. (2012). The impact of coffee certification on small-scale producers' livelihoods: A case study from the Jimma Zone, Ethiopia. *Agricultural Economics*, 43(4): 429–440. https://doi.org/10.1111/j.1574-0862.2012.00594.x

Jensen, R. and Oster, E. (2009). The power of TV: Cable television and women's status in India. *The Quarterly Journal of Economics*. 124(3): 1057–1094. https://www.jstor.org/stable/40506252?seq=1#page_scan_tab_contents

Khandker, S., Koolwal, G. and Samad, H. (2010). *Handbook on impact evaluation: Quantitative methods and practices*. Washington, DC: World Bank.

Kilic, T., Zezza, A., Carletto, C. and Savastano, S. (2017). Missing(ness) in action: selectivity bias in GPS-based land area measurements. *World Development*, 92: 143–157. https://doi.org/10.1016/j.worlddev.2016.11.018

King, G., Nielson, R., Coberley, C., Pope, J. E. and Wells, A. (2011). Avoiding randomization failure in program evaluation, with application to the Medicare Health Support Program. *Population Health Management*, 14(1): S11–S22. https://dash.harvard.edu/bitstream/handle/1/5125263/mhs.pdf?sequence=1

Kling, J. R., Liebman, J. B. and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1): 83–119. https://doi.org/10.1111/j.1468-0262.2007.00733.x

Korte, J. E., Djimeu, E. W. and Calvo, F. A. (2018). Evidence of behavioural compensation in internal replication of male circumcision trial to reduce HIV acquisition in Kisumu, Kenya. Manuscript accepted for publication, *Journal of Development Studies*.

Kuecken, M. and Valfort, M. (2018). Information reduces corruption and improves enrolment (but not schooling): A replication study of a newspaper campaign in Uganda. Manuscript accepted for publication, *Journal of Development Studies.*

Lakens, D. (2016). Why you don't need to adjust your alpha level for all tests you'll do in your lifetime. The 20% Statistician. Available at: http://daniellakens.blogspot.ch/2016/02/why-you-dont-need-to-adjust-you-alpha.html

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, 24(4): 414–433. https://econpapers.repec.org/article/cuppolals/v_3a24_3ay_3a2016_3ai_3a04_3ap_3a414-433_5f01.htm

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4): 604–620. https://www.jstor.org/stable/1806062?seq=1#page_scan_tab_contents

Lampach, N. and Morawetz, U. B. (2016). Credibility of propensity score matching estimates. an example from Fair Trade certification of coffee producers. *Applied Economics*, 48(44): 4227–4237. https://doi.org/10.1080/00036846.2016.1153795

Lee, D. S., Moretti, E. and Butler, M. J. (2004). Do voters affect or elect policies? Evidence from the US House. *The Quarterly Journal of Economics*, 119(3): 807–859. https://doi.org/10.1162/0033553041502153

Lim, S. S., Dandona, L., Hoisington, J. A., James, S. L., Hogan, M. C. and Gakidou, E. (2010). India's Janani Suraksha Yojana, a conditional cash transfer programme to increase births in health facilities: An impact evaluation. *The Lancet*, 375(9730): 2009–2023. https://doi.org/10.1016/S0140-6736(10)60744-1

Malesky, E., Schuler, P. and Tran, A. (2012). The adverse effects of sunshine: a field experiment on legislative transparency in an authoritarian assembly. *The American Political Science Review*, 106(4): 762–786. https://www.jstor.org/stable/23357708?seq=1#page_scan_tab_contents

Matern, R., Mendelson, M. and Oliphant, M. (2009). Testing the validity of the Ontario deprivation index. Daily Bread Food Bank and the Caledon Institute of Social Policy. https://maytree.com/wp-content/uploads/837ENG.pdf

McKenzie, D. (2017). Should we require balance t-tests of baseline observables in randomized experiments. Development Impact. Available at: http://blogs.worldbank.org/impactevaluations/should-we-require-balance-t-tests-baseline-observables-randomized-experiments

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1): 159–217. https://doi.org/10.1111/j.1468-0262.2004.00481.x

Moundigbaye, M., Rea, W. S. and Reed, W. R. (2018). Which panel data estimator should I use? A corrigendum and extension. *Economics: The Open-Access, Open-Assessment E-Journal*, 12(2018-4): 1– 31. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-4

Nunn, N. and Qian, N. (2014). US food aid and civil conflict. *AmericanEconomic Review*. 104(6): 1630-1666. https://www.aeaweb.org/articles?id=10.1257/aer.104.6.1630

Ozler, B. (2014). Obesity may not have dropped among children, but it almost certainly increased among the elderly. Development Impact. Available at: http://blogs.worldbank.org/impactevaluations/obesity-may-not-have-dropped-among-children-it-almost-certainly-increased-among-elderly

Parada, J. (2016). Access to modern markets and the impacts of rural road rehabilitation: evidence from Nicaragua. University of California. https://arefiles.ucdavis.edu/uploads/filer_public/04/0a/040a8605-163d-4888-b6af-0dd0b61c910f/parada_jmp.pdf

Powell-Jackson, T., Davey, C., Masset, E., Krishnaratne, S., Hayes, R., Hanson, K. and Hargreaves, J. R. (2018). Trials and tribulations: cross-learning from the practices of epidemiologists and economists in the evaluation of public health interventions. *Health Policy and Planning*, 33(5): 702–706. https://doi.org/10.1093/heapol/czy028

Reed, W. R. (2017). Replication in labor economics. *IZA World of Labor 2018: 413.* https://wol.iza.org/articles/replication-in-labor-economics

Reinikka, R. and Svensson, J. (2005). Fighting corruption to improve schooling: evidence from a newspaper campaign in Uganda. *Journal of the European Economic Association*, 3(2-3): 259−267. https://www.researchgate.net/publication/24090693_Fighting_Corruption_to_Improve_Schooling_Evidence_from_a_Newspaper_Campaign_in_Uganda

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4): 1237−1282. https://doi.org/10.1111/j.1468-0262.2005.00615.x

Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, 71(1): 135–158. https://doi.org/10.1111/j.1468-0084.2008.00542.x

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125 (1): 175–214. https://ideas.repec.org/a/oup/qjecon/v125y2010i1p175-214..html

Rothstein, J. (2017). Measuring the impacts of teachers: comment. *American Economic Review*, 107(6): 1656-1684. https://pubs.aeaweb.org/doi/pdf/10.1257/aer.20141440

Samii, C. (2016). Inverse covariance weighting versus factor analysis. Cyrus Samii blog. Available at: http://cyrussamii.com/?p=2177

Scherer, T. L. (2015). The OECD's fragility index is surprisingly fragile and difficult to reproduce. *The Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2015/05/17/the-oecds-fragility-index-is-surprisingly-fragile-and-difficult-to-reproduce/?utm_term=.af4cbf192c60

Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1–2): 305–353. https://doi.org/10.1016/j.jeconom.2004.04.011

Stuart, E. A. (2010). Matching methods for causal inference: a review and look forward. *Statistical Science*, 25(1): 1–21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943670/

Varadhan, R. and Seeger, J. D. (2013). Chapter 3: Estimation and reporting of heterogeneity of treatment effects. In Velentgas, P., Dreyer N. A., Nourjah, P., Smith, S. R. and Torchia, M. M. (Eds.), *Developing a protocol for observational comparative effectiveness research: a user's guide*. Rockville, MD: Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK126188/

Waddington, H., Aloe, A.M., Becker, B. J., Djimeu, E. W., Hombrados, J. G., Tugwell, P., Wells, G. and Reeves, B. (2017). Quasi-experimental designs series – Paper 6: risk of bias assessment, *Journal of Clinical Epidemiology*, 89: 43–52. https://doi.org/10.1016/j.jclinepi.2017.02.015

Wood, B.D.K. and Brown, A.N. (2015). What 3ie is doing in the replication business. The Replication Network. Available at: https://replicationnetwork.com/2015/10/15/benjamin-wood-and-annette-brown-what-3ie-is-doing-in-the-replication-business/

Wood, B.D.K. and Dong, M. (2015). Recalling extra data: a replication study of 'finding missing markets'. Manuscript accepted for publication, *Journal of Development Studies.*

Wood, B. D. K., Müller, R. and Brown, A. N. (2018). Push button replication: Is impact evaluation evidence for international development verifiable? Manuscript submitted for publication.

Young, A. (2017). Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. Available at: http://personal.lse.ac.uk/YoungA/ChannellingFisher.pdf

Zimmerman, C. (2015). On the need for a replication journal. Working paper 2015-016A. Federal Reserve Bank of St. Louis. https://pdfs.semanticscholar.org/d31d/8d9d92250ee8ccb6d7f10b9f2c7efbbb832c.pdf