Brown, Annette N.; Wood, Benjamin Douglas Kuflick

Working Paper

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

Economics Discussion Papers, No. 2017-77

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 (2017) : Which tests not witch hunts: a diagnostic approach for conducting replication research, Economics Discussion Papers, No. 2017-77, Kiel Institute for the World Economy (IfW), Kiel

This Version is available at:
https://hdl.handle.net/10419/169136

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

Discussion Paper
No. 2017-77 | September 28, 2017 | http://www.economics-ejournal.org/economics/discussionpapers/2017-77

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

Annette N. Brown and Benjamin Douglas Kuflick Wood

Abstract
This paper provides researchers with an objective list of checks to consider when planning a replication study with the objective of validating findings for informing policy. These replication studies should begin with a pure replication of the published results and then reanalyse the original data to address the original research question. The authors present tips for replication exercises in four categories: validity of assumptions, data transformations, estimation methods, and heterogeneous impacts. For each category they offer an introduction, a tips checklist, some examples of how these checks have been employed, and a set of resources that provide statistical and econometric details.

(Published in Special Issue The practice of replication)

JEL C10 B41 A20
Keywords Replication; diagnostic; validation; impact evaluation; reanalysis; risk of bias

Authors
Annette N. Brown, FHI 360
Benjamin Douglas Kuflick Wood, ✉ International Initiative for Impact Evaluation, Washington Office, USA, bwood@3ieimpact.org

Citation Annette N. Brown and Benjamin Douglas Kuflick Wood (2017). Which tests not witch hunts: a diagnostic approach for conducting replication research. Economics Discussion Papers, No 2017-77, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2017-77

Received September 8, 2017 **Accepted as Economics Discussion Paper** September 12, 2017
Published September 28, 2017

1. Introduction

While most researchers accept the scientific premise for replication, many still oppose or resist its practice. One reason given for this resistance is the assumed intent of replication research, especially of *internal* replication research where the replication researcher works with the data from the original study. Galiani, Gertler and Romero (2017) claim that replication suffers from "overturn bias", which they attribute to both journal editors and to replication researchers, citing a survey of editors as evidence of the first claim and original authors of replicated studies as evidence of the second claim. 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. One challenge for avoiding "overturn bias" and using replication to strengthen positive findings is to design the right replication plan. When looking for tests to run, researchers instinctively look for what seems "wrong". In this paper, we present 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".

We were motivated by the feedback we received from the replication researchers we worked with under the International Initiative for Impact Evaluation's (3ie's) Replication Program.[1] 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 determined 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 document from the original study.

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 robustness checks, however, as many of the threats to bias identified in risk-of-bias tools concern things that cannot be changed once the data are collected, such as whether treatment status is blinded.

In compiling the 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 examples in four categories: validity of assumptions, data transformations, estimation methods, and heterogeneous impacts. The first category includes

[1] Annette Brown directed the 3ie Replication Program from its establishment through July 2016. Benjamin Wood has managed the program from 2012 to the present.

checks on the setup of the study and the use of estimation methods. The tests do not re-analyze the data to answer the original study's research question, they merely use the data to check the assumptions behind the analysis as conducted in the original.

The middle two categories relate to the measurement and estimation analysis approach to replication (Brown, Cameron and Wood, 2014) which encompasses much of the analysis we see in replication studies. Sukhtankar (2017) catalogs a set of published replication studies in development economics and finds that "the majority of replications involve re-analyses using difference econometric specifications or reconfigurations of the data." (p. 33) The fourth category, heterogeneous outcomes, is related to the theory of change analysis approach to replication, as the question of whether the results are different for different subgroups is a straightforward check on the theory and how it might be applied more generally.

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 validating results. Ideally both theory and context inform the exercises chosen by replication researchers in all four categories. For example, checking balance for additional variables should be done for variables that are predicted to be influential. Alternate variable constructions should be motivated by a careful understanding of the concept and the context. And so on.

For each of the four categories, we provide an overview, list some suggested tips, and then give examples from 3ie-funded replication research.[2] We do not provide detailed statistical descriptions of suggested robustness checks. Rather we suggested some selected resources at the end of each section.

2. Validity of assumptions

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). Researchers often use balance tests to assess group equivalence, but sometimes visual comparisons and other tests can also be useful, especially if a replication researcher considers the similarity of distributions between the treatment and comparison to be pertinent. 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.

[2] The references are currently to the 3ie Replication Paper Series versions of the studies, except for Davey et al. (2015). It is expected that most of the studies will be published in an upcoming journal supplement.

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.

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 (2017) 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.

Tips for exercises to validate assumptions

- ☐ Test balance for additional, relevant variables
- ☐ Test balance at applicable units of analysis or for analyzed subsets of the data
- ☐ 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 across time
- ☐ Identify important untested assumptions for chosen estimation methods and test using accepted methods

2.1. Examples

In their replication study of Galiani and Schargrodsky's impact evaluation of property rights for the poor, Cameron, Whitney, and Winters (2015) recognize that the balance tests reported in the original paper 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. report 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.

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.

Resources for validating assumptions

Citation	Key words
Bärnighausen et al. (2017) "Quasi-experimental study designs series – Paper 7: assessing the assumptions"	Quasi-experimental designs, instrumental variables, regression discontinuity, interrupted time series, fixed effects, difference-in-differences, assumption tests, monotonicity, continuity
Bruhn and McKenzie (2009) "In pursuit of balance: randomization in practice in development field experiments"	Balance tests
McKenzie (2017) "Should we require balance t-tests of baseline observables in randomized experiments?"	Balance tests, sample attrition, randomization implementation
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

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. Thomas Scherer (2015) replicates the OECD fragility index and finds that more than half of the countries measured by the OECD are misclassified.

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

3.1. Examples

In the Basurto, Burga, Flor Toro, and Huaroto (2015) replication study, the researchers note that the development of the HHH2009 dataset used by Cattaneo, Galiani, Gertler, Martinez, and Titiunik (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 programmes*, *total per capita value of household assets*, and *total per capita consumption*, Arithmetic Mean Imputation 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, Burga, Flor Toro, and Huaroto (2015) include an appendix reviewing the literature around imputation methods.

Korte, Djimeu 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. Korte et al. conclude that the original results are not sensitive to missing data.

Kuecken and Valfort's (2015) 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 argue that such shocks should not be systematically correlated with the explanatory variable. After reintegrating the dropped schools into the sample, 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 (2014) 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 analysz 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.

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. 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. Imlach, Gunesekara, Carter and Blakely (2008) provide a glossary to help epidemiologists and econometricians understand each other's languages. A well-known example of an epidemiological replication study of an econometric paper is Davey, Aiken, Hayes, and Hargreaves (2015). We provide 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. 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. Another example of different approaches to a quasi-experimental estimation method is regression discontinuity design. Button (2015), 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. Lakens (2016) provides a useful discussion of the latter.

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 analyse 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

4.1 Examples

In the Korte et al. (2015) 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. 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, Djimeu and Calvo's pure replication, thus providing strong support for the main findings of the original paper.

Cameron et al. (2015) 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 (2015) 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) 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.

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, M (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, B (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, J and Wolf, M (2005) "Stepwise Multiple Testing as Formalized Data Snooping"	Bootstrap, data snooping, familywise error, multiple testing, stepwise method.
Young, A (2016) "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, J and Coopers, JJ (n.d.) "10 things to know about cluster randomization"	Information reduction, cluster sizes, within-cluster spillovers, power analysis

5. Heterogeneous impacts

The replication study by Cervellati, Jung, Sunde, and Vischer (2014) of Acemoglu, Johnson, Robinson, and Yared'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 impacts. One is simply statistical power. If the dataset is small to begin, there may not be enough power to meaningfully conduct subgroup analysis. 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. And finally, subgroup analysis introduces the multiple comparison, or multiple hypothesis tests, issue discussed in the previous section. These are all important considerations for replication researchers who explore heterogeneous impacts.

Tips for heterogeneous impacts exercises

- ☐ Identify theoretically or clinically relevant subgroups and check whether heterogeneous impacts are tested for these subgroups
- ☐ Test for heterogenous impacts for relevant subgroups
- ☐ Search for variation in treatment effects using machine learning methods

5.1 Examples

Carvalho and Rokicki (2015) 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 further 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 by previous export crop producer status, focusing specifically on the three crops promoted in the intervention. Based on value-chain theory, the replication researchers suspected previous participation in markets might yield value information on this type of intervention, and explored this possibility through this theory of change reanalysis. The original heterogeneous impacts results proved robust to this alternative approach.

Iverson and Palmer-Jones (2014) provide another example of exploring alternatives theory of change through replication research. 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 attempt to unpack the theory of change by examining the mechanisms leading to the change in the observed outcomes in more detail. Ultimately, the heterogeneous outcomes reanalysis suggests that cable TV access may influence certain women's rights more than others, for example 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.

Resources for heterogenous impacts

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, R and Seeger, JD (2013) "Chapter 3: Estimation and reporting of heterogeneity of treatment effects"	Heterogeneity of treatment effect, clinically relevant subgroups, observational comparative effectiveness research
Cummins, JR (2017) "Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment"	Outcome distribution, targeting of behavioural responses, rank similarity, generalizability
Athey, S and Imbens, GW (2017) "The econometrics of randomized experiments"	General treatment effect heterogeneity, covariates, valid confidence intervals

6. Conclusion

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 with the objective of validating research findings for informing policy. Replication studies with this objective should begin with a pure replication of the published results but then extend beyond the pure replication to reanalyze the original data to address the original research question. We present the replication exercises in four categories: validity of assumptions, data transformations, estimation methods, and heterogeneous impacts. For each category we offer an introduction to the issues, some examples of how these checks are employed across a collection of replication studies of development impact evaluations, and a set of resources that provide statistical and econometric details.

References

Acemoglu, D., Johnson, S., Robinson, J. A., & Yared, P. (2008). Income and democracy. *American Economic Review,* 98(3), pp. 808–842.

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

Alsop, R., Pertelsen, M., and Holland, J. (2006). *Empowerment in practice: analysis to implementation, Directions in Development*. Washington, DC: World Bank.

Anderson, M. (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.

Ashraf, N., Giné, X. & 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, pp. 973–990.

Athey, S. & Imbens, G. W. (2017). Chapter 3 – The econometrics of randomized experiments. In A. Banerjee and E. Duflo (Eds.) *Handbook of Economic Field Experiments* (pp. 73-140). Elsevier.

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

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. & Vollmer, S. (2017). Quasi-experimental study designs series – Paper 7: assessing the assumptions, *Journal of Clinical Epidemiology*, doi: 10.1016/j.jclinepi.2017.02.017.

Basurto, M. P., Burga, R., Flor Toro, J. L. & Huaroto, C., (2015). Walking on solid ground: a replication study on Piso Firme's impact, 3ie Replication Paper 7. Washington, DC: International Initiative for Impact Evaluation (3ie).

Bowers, J. & Cooper, J. J. (n.d.). 10 things to know about cluster randomization. *Evidence in Governance and Politics Methods Guides*, [online]. Available at: http://www.egap.org/methods-guides/10-things-you-need-know-about-cluster-randomization [Accessed June 7, 2017].

Bowser, W. (2015). The impact of agricultural extension and roads on poverty and consumption growth in fifteen Ethiopian villages: a replication study, 3ie Replication Paper 4. Washington, DC: International Initiative for Impact Evaluation (3ie).

Brown, A. N., Cameron, D. B. & 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, DOI: 10.1080/19439342.2014.944555

Bruhn, M. & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments, *American Economic Journal: Applied Economics*, 1:4, 200-232.

Button, P. (2015). *A replication of 'Do voters affect or elect policies? Evidence from the US house' (Quarterly Journal of Economics, 2004)* (Tulane Economics Working Paper Series No. 1518). Retrieved from Tulane University website: http://econ.tulane.edu/RePEc/pdf/tul1518.pdf.

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. Washington, DC: International Initiative for Impact Evaluation (3ie).

Carvalho, N. & Rokicki, S. (2015). The impact of India's JSY conditional cash transfer programme: A replication study, 3ie Replication Paper 6. Washington, DC: International Initiative for Impact Evaluation (3ie).

Cattaneo, M. D., Galiani, S., Gertler, P. J., Martinez, S. & Titiunik, R. (2009). Housing, health, and happiness. *American Economic Journal: Economic Policy*, 1(1), pp.75–105.

Ceasar de Andrade, S., Previdelli, Á., Lobo Marchioni, D. & Fisberg, R. (2013). Evaluation of the reliability and validity of the Brazilian Healthy Eating Index Revised. *Revista de Saúde Pública*, 47(4), pp. 1-7.

Cervellati, M., Jung, F., Sunde, U. & Vischer, T. (2014). Income and democracy: comment. *American Economic Review*, (2014), 104(2), pp. 707-719.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2017). Measuring the impacts of teachers: reply. *American Economic Review,* (107)6, pp. 1685-1717.
Deaton, A. (2010)

Cummins, J. R. (2017). Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment. *Economics of Education Review*, 56, pp. 40-51.

Davey, C., Aiken, A. M., Hayes, R. J. & 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), pp. 1581-1592.

De la Cuesta, B. & Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections, *Annual Review of Political Science*, 19, pp. 375-396.

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.

Djimeu, EW, Korte JE and Calvo, FA. (2015). Male circumcision and HIV acquisition: reinvestigating the evidence from young men in Kisumu, Kenya, 3ie Replication Paper 8. Washington, DC: International Initiative for Impact Evaluation (3ie).

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association,* 56 (293), pp. 52-64.

Evidence in Governance and Politics. (n.d.) 10 things to know about heterogeneous treatment effects. *Evidence in Governance and Politics Methods Guides,* [online]. Available at: http://www.egap.org/content/10-things-know-about-heterogeneous-treatment-effects [Accessed June 7, 2017].

Galiani, S., Gertler, P. & Romero, M. (2017) Incentives for replication in economics. Working Paper 23576. Cambridge, MA: National Bureau of Economic Research.

Galiani, S. & Schargrodsky, E. (2010). Property rights for the poor: effects of land titling. *Journal of Public Economics*, 94(9–10), pp. 700–79.

Gertler, P., Martinez, S., Premand, P., Rawlings, R. & Vermeersch, C. (2016). *Impact Evaluation in Practice*. Second edition. Washington, DC: The International Bank for Reconstruction and Development.

Goel, R. & 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), pp. 143-154.

Helland, E. & Tabarrok, A. (2004). Using placebo laws to test “more guns, less crime”. *The B.E. Journal of Economic Policy and Analysis,* 4(1).

Imai, K. & Ratkovic, M. (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1), pp. 443-470.

Imai, K. & 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, pp. 1-19.

Imbens, G. W. & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies,* 79(3), pp. 933-959.

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature,* 47(1), pp. 5-86.

Imlach Gunasekara, F., Carter, K. & Blakely, T. (2008). Glossary for econometrics and epidemiology. *Journal of Epidemiological Community Health,* 62, pp. 858-861.

Iversen, V. & Palmer-Jones, R. (2014). TV, female empowerment and demographic change in rural India, 3ie Replication Paper 2. Washington, DC: International Initiative for Impact Evaluation (3ie).

Jensen, R. & Oster, E. (2009). The power of TV: Cable television and women's status in India. *The Quarterly Journal of Economics*. 124(3), pp. 1057-1094.

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

Kilic, T., Zezza, A., Carletto, C. & Savastano, S. (2017). Missing(ness) in action: selectivity bias in GPS-based land area measurements. *World Development,* 92, pp. 143–157.

King, G., Nielson, R., Coberley, C., Pope, J.E. & Wells, A. (2011). Avoiding randomization failure in program evaluation, with application to the Medicare Health Support Program. *Population Health Management,* 14(1), pp. S11-S22.

Kling, J., Liebman, J. & Katz, L. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1), pp. 83–119.

Kuecken, M. & Valfort, M. (2015). Fighting corruption does improve schooling: a replication study of a newspaper campaign in Uganda, 3ie Replication Paper 10. Washington, DC: International Initiative for Impact Evaluation (3ie).

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*, [online]. Available at: http://daniellakens.blogspot.ch/2016/02/why-you-dont-need-to-adjust-you-alpha.html [Accessed June 6, 2017].

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis,* 24(4), pp. 4414-433.

LaLonde, R. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, 76(4), pp. 604-620.

Lee, D. S., Moretti, E. & Butler, M. J. (2004). Do Voters Affect or Elect Policies? Evidence from the US House. *Quarterly Journal of Economics*, 119(3), pp. 807-59

Lim, S. S., Dandona, L., Hoisington, J. A., James, S. L., Hogan, M. C. & 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), pp. 2009–2023.

Matern, M., Mendelson, M. & Oliphant, M. (2009). *Testing the Validity of the Ontario Deprivation Index*. Daily Bread Food Bank and the Caledon Institute of Social Policy. Toronto, ON: Metcalf Foundation.

McKenzie, D. (2017). How Effective Are Active Labor Market Policies in Developing Countries?

A Critical Review of Recent Evidence, Policy Research Working Paper 8011. Washington, DC: The World Bank.

McKenzie, D. (2017). Should we require balance t-tests of baseline observables in randomized experiments, World Bank Development impact, [online]. Available at: http://blogs.worldbank.org/impactevaluations/should-we-require-balance-t-tests-baseline-observables-randomized-experiments [Accessed 7 September 2017].

Ozler, B. (2014). Obesity may not have dropped among children, but it almost certainly increased among the elderly, World Bank Development Impact, [online]. Available at: http://blogs.worldbank.org/impactevaluations/obesity-may-not-have-dropped-among-children-it-almost-certainly-increased-among-elderly [Accessed 8 April 2017].

Parada, J. (2017). Access to modern markets and the impacts of rural road rehabilitation: Evidence from Nicaragua. University of California, Davis job market paper.

Reinikka, R. & Svensson, J. (2005). Fighting corruption to improve schooling: evidence from a newspaper campaign in Uganda. *Journal of the European Economic Association*, 3(2-3), pp. 259–67.

Romano, J. & Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica*, 73(4), pp. 1237-1282.

Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, 71 (1), pp. 135–158.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *Quarterly Journal of Economics,* 125 (1), pp. 175-214.

Rothstein, J. (2017). Measuring the impacts of teachers: comment. *American Economic Review,* 107(6), pp. 1656-1684.

Samii, C. (2016). Inverse covariance weighting versus factor analysis. Cyrus Samii blog. [online]. Available at: http://cyrussamii.com/?p=2177 [Accessed 5 June 2017].

Scherer, T.L. (2015). The OECD's fragility index is surprisingly fragile and difficult to reproduce, New York Times Monkey Cage [online]. Available at: 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 [Accessed 7 September 2017].

Schultz, K., Altman, D. & Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *The Lancet*, 357(9263), pp. 1191–1194.

Smith, J. A. & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics,* 125, pp. 305-353.

Stuart, E. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science,* 25(1), pp. 1-21.

Sukhtankar, S. (2017). Replications in development economics. *American Economic Review: Papers & Proceedings 2017*, 107(5), pp. 32-36.

Varadhan, R. & Seeger, J. D. (2013). Chapter 3: Estimating and reporting of heterogeneity of treatment effects. In P. Velentgas, N. A. Dreyer, P. Nourjah, Smith S. & Torchia, M. M. (Eds.), *Developing a protocol for observational comparative effectiveness research: a user's guide* (pp. 35-44). Rockville, MD: Agency for Healthcare Research and Quality.

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

Whiting, P., Savović, J., Higgins, J., Caldwell, D., Reeves, B., Shea, B., Davies, P., Kleijnen, J., Churchill, R. & ROBIS group. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, 69, pp. 225–234.

Wood, B.D.K. & Brown, A.N. (2015) What 3ie is doing in the replication business. October 15. In: Guest Blogs. 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. & Dong, M. (2015). Recalling extra data: a replication study of Finding missing markets, 3ie Replication Paper 5. Washington, DC: International Initiative for Impact Evaluation (3ie).

Young, A. (2016). Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results, [online]. Available at: http://personal.lse.ac.uk/YoungA/ChannellingFisher.pdf [Accessed 8 April 2017].