This document distills “Understanding Replication, A Historical and Conceptual Guide” by Vegard Iversen and Richard Palmer-Jones (IPJ) for researchers planning internal replications of previous impact evaluations. Parts of the following are taken verbatim from IPJ’s current working paper and the 3ie Replication Program Document. Please note that this is the June 2012 version of the practical guidelines. 3ie plans to expand and update the guidelines over time based on the experiences of the replication program. Send any questions or comments regarding these guidelines to replication@3ieimpact.org.

3ie’s Practical Guidelines to Replication of Impact Evaluations

The mission of the 3ie replication program is to improve the quality of evidence available for development policy making and program design by increasing the conduct of internal replication studies of impact evaluations of development programs. By internal replication, we mean the re-analysis of the original data and, in some cases, the validation of the data analysis by using different data sources for the same population. The 3ie program does not include external replication—replications of interventions in different settings to assess the external validity of an impact evaluation of a particular intervention. This document is designed to assist replication researchers to understand 3ie’s terminology and standards pertaining to replications.

As stated in 3ie’s Replication Program Document, all proposed replications must be of an impact evaluation listed on the replication program’s Candidate Studies List. To be eligible for the list, study must constitute an impact evaluation according to 3ie’s Principles of Impact Evaluation, including demonstration of attributable change from a program or intervention using a recognized impact evaluation estimation technique. Studies selected for the list have been identified as innovative, influential, or controversial by the development community and 3ie’s Replication Program Advisory Group. The list is dynamic; as we fund or learn about replications of studies on the list, we will add new studies to the list. Suggestions for the list are always welcome.

Replications can suffer from the same kinds of data mining and reporting biases as primary studies. To mitigate these problems, 3ie will publish the replication plans for all 3ie-funded replications on our website. The replication plans published will be the “final” versions, after replication researchers have had an opportunity to revise the plans from their applications according to reviewer comments. The revision process will occur during the data acquisition period so that final replication plans will be published prior to the initiation of the analysis.

Categories of replication

There exist different categorizations of replication in social science. Drawing on the work of Hamermesh (2007) and IPJ, 3ie classifies internal replication research into three categories: pure,
statistical, and scientific. Pure replication is the reproduction and reconciliation of the results published in the original study using the same data and techniques. Pure replication requires more than re-running the original programming files on the cleaned data. It involves independent construction of the variables from the raw data and re-estimation using the study’s methodologies. This reconciliation process, if successful, validates the study’s original findings, which is the first step to providing credible evidence for policymakers.

To strengthen the credibility of results, the next two categories of replication—statistical and scientific—move beyond reconciliation to examine the robustness of the findings to different statistical methods, different data, or alternate theories of change. Statistical replication is the re-analysis of the original study’s hypotheses using different data treatments (i.e. different variable constructs or statistical techniques) and may include a comparison of the study’s data to other available data. Some examples of statistical replication exercises are: using the original data to construct key variables in different ways (e.g. using highest degree earned instead of years of education to measure education); testing different treatments of outliers for key variables; comparing data from the project survey to data from national surveys; and using different estimation techniques to test the same hypotheses.

Scientific replication seeks to validate the findings of the original study by using the study’s own data to further explore the hypothesized theory of change and/or test alternate theories of change. While pure replication may show that the original results are valid (no errors were made) and statistical replication may show that the original results are robust to different variable constructs and estimation methodologies, the usefulness of the results for policy makers often still depends on whether the results truly reveal how the intervention works (or does not work)—that is, whether the results demonstrate what theory of change is at work. Some examples of scientific replication exercises are: testing for heterogeneous outcomes not done in the original study and estimating different specifications based on alternate hypotheses. Note that scientific replication is not a data mining exercise where any and all possible specifications are tested. Rather scientific replication begins with proposed alternate theories of change, which are identified before the testing begins.

The line between scientific replication and extensions to earlier research is a fuzzy one. Often research that is intended to build on a prior study begins with what is essentially scientific replication of the preceding study. One distinguishing factor in the 3ie terminology is that all replications must begin with pure replication. That is, a replication starts at the beginning of the subject study in order to first reconcile the original findings, but may include extensions to the study with new findings perhaps supporting different theories of change. In contrast, a non-replication extension study begins near the end of the subject study and focuses on doing something new and different. Table 1 provides a summary of the three categories of replication.
Table 1 Objectives and characteristics of replications (IPJ, Hamermesh, 2007)

<table>
<thead>
<tr>
<th>Pure</th>
<th>Statistical</th>
<th>Scientific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconciliation of the original empirical results</td>
<td>(Pure +) Different data source and/or Different variable concepts and measurement and/or Different estimation methods</td>
<td>(Pure +) Deliberate introduction of alternative conceptual framework and methods</td>
</tr>
</tbody>
</table>

**Steps in pure replication**

3ie recommends the following steps for all pure replications:

- Reconciliation of the original and the recalculated descriptive statistics of the raw data for all primary variables;
- Reconciliation of the descriptive statistics for all constructed variables with the reconstructed variables from the raw data and reconciliation of the samples used for estimation with a reconstruction of the samples from the raw data;
- Reconciliation of estimation results in the original study with the re-estimation of the original specifications.

The first step examines the code for data capture, cleaning, and descriptive statistics, including side by side comparison of descriptives reported in the original study and the replication including sample size, mean, and standard deviation, at least. The second step examines the code to create final variables and final samples used in the analysis, including side-by-side tables of descriptives in the original study and the replication. The final step involves re-writing the estimation code and conducting a side-by side comparison of results in the original study and from the re-estimation.

In order to conduct pure replication, the replication researcher begins by obtaining the original data, programming files, and accompanying survey materials. The researcher then independently programs the original analysis and reconciles those results with the previous findings. It is particularly important to note differences in sample sizes, outlier assumptions, or the use of weights. Descriptive statistic comparisons oftentimes play a helpful role in reconciling differences in findings.

Replication researchers should draw attention to discrepancies and report on efforts at reconciliation. At each step, the research should record any steps that had to be taken beyond reading the original study to accomplish the replication. Reprogramming in different software, or entirely independently, is a way to reduce the effects of assumptions built into the software.
and/or identify issues with the author’s programming. An alternative to using different software is to use different methods in the same software to capture, clean and compute variables.

If the original study was publicly registered, the pure replication should include a reconciliation of how the original research was conducted against the research plan recorded in the registry.

Pure replication on its own is a rather limited way to improve the credibility of evidence used for policy making. More is learned, and more confidence can be warranted, if findings are supported by multiple methods. Underlying the need for corroboration is the realization of how everyday orthodoxies built on apparently robust analysis can be overthrown by new knowledge. Statistical and scientific replication provide additional corroboration.

*Elements of statistical and scientific replication*

As noted above, all 3ie-funded replication studies must begin with pure replication. Not only does pure replication check for errors in the original research, the process of pure replication ensures that the replication researcher understands the original data and analysis and helps her/him to confirm whether the proposed statistical and/or scientific replication is well conceived.

Statistical and scientific replications explore different estimation technique or test alternative theories theoretical frames or theories. These replications are often warranted as most social science is characterized by controversy and strongly contrasting accounts of phenomena (Lakatos and Musgrave, 1970). Statistical or scientific replications may include:

- Comparison of descriptive statistics with those for equivalent populations from other credible sources (alternative data-sets);
- Reconstruction of variables based on different measurement practices or conceptions of how particular phenomena should be measured;
- Estimation on different samples of the data reflecting different treatment of outliers or testing for heterogeneous outcomes;
- Estimation using different statistical or econometric techniques drawing on “best practices” in social science research;
- Testing alternative hypotheses or theories of change after a thorough review of the literature to identify alternative approaches or framings of the issues at stake.

Particularly for scientific replication, it is important that both the justification and specifications for the new estimations be set out in advance. These are elements of the replication plan. For 3ie-funded replications, these plans are published online.

---

1 This is well known in the medical world and helped inspire the systematic review movement, as documented for example by Goldacre (2008).
Good replication practice includes testing alternative credible theories, since testing only the maintained hypothesis is not sufficient to provide robust support for the main conclusions drawn in the original study, particularly if these conclusions are intended to influence policy making and program design. Iversen and Palmer-Jones (2008) make this argument in relation to Basu et al. (2002). We may also draw an analogy with the arguments in the bio-medical arena that bias is introduced if pharmaceuticals are tested only against placebos rather than the standard treatment, or the alternatives provided by competitors. This is because it may not be difficult to find an effect, but what is necessary knowledge for patient treatment is whether the net effect (all benefits net of undesirable side effects) is greater than all alternatives.

**Reporting**

It is important that the replication researcher maintains open communication with the original authors. Particularly during the pure replication stage, the replication researcher should make all possible attempts to obtain the information needed from the authors in order to reconcile the original findings. The replication researcher should also inquire as to whether the exercises planned for statistical replication were conducted but not reported or just not conducted. All communications with the original authors should be documented.

A pure replication that fully reconciles the statistics and results published in the original paper is considered a “successful” pure replication. Sometimes pure replications may only be partly successful because the raw data are no longer available. In these cases, the pure replication exercises that can be conducted may be successful but others may not be possible. A successful pure replication in no way precludes statistical or scientific replication, as these stages explore different questions. In cases where the pure replication is successful, these results should be reported in the final replication study even though the study may go on to present findings from statistical and/or scientific replication.
Bibliography


