Achieving robustness in the E&A macro evaluation: A Technical Note

Introduction

In developing and piloting a methodology for the E&A macro evaluation, we have identified a number of issues around achieving robustness in the evaluation research methodology. In this Note we pull these issues together around the following three robustness principles and one cross cutting principle. The three robustness principles are:

- The first principle of **reliability** ensures that a result achieved with the chosen research method can be repeated by any given researcher. Reliability builds confidence in the repeatability of a study's given research method;
- The second principle of **internal validity** is applied to studies that attempt to establish a causal relationship. It allows us to be confident that a changing project outcome can be attributed to a given intervention. Internal validity builds confidence in the cause and effect relationship within a study;
- The third principle of **external validity** increases our confidence that we can generalise results beyond the immediate study population, thus building 'confidence of inference' from that study.

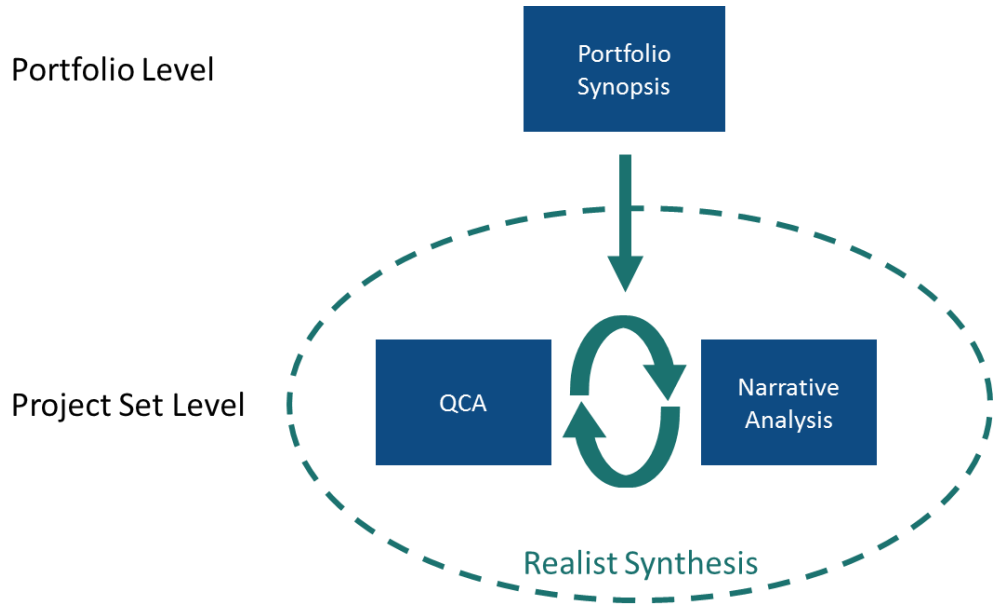
Cross cutting these three principles is a fourth principle of **transparency**. This requires that the application of these robustness principles through research protocols is open to external scrutiny by third parties, enabling challenge and verification.

Applying these principles in practice is strongly influenced by the type of research methodology employed. Standard experimental, empiricist research bring with it a clear set of procedures for increasing the reliability and (internal and external) validity of study. We have adapted these robustness principles to the application of our chosen realist synthesis[1] research approach for the macro evaluation (see Figure 1).[2] Rather than seeking universal truths based on inflexible methods, a realist synthesis seeks to negotiate the complexities of programme interventions by identifying underlying causal mechanisms and exploring how they work in particular contexts and settings.[3]

Our approach sequences a pattern-finding **QCA method** that identifies significant 'causal configurations' of factors (or conditions) that are associated with a given project outcome, with an interpretive **narrative analysis method** that examines these causal configurations in greater depth and explores how they work in different contexts and under what conditions.

[1]See Pawson, R., Greenhalgh, T., Harvey, G. and Walshe, K. (2004),'Realist synthesis: an introduction', *RMP Methods Paper* 2/2004. Manchester, UK: ESRC Research Methods Programme, University of Manchester.
[2] For a fuller discussion of the methodology, see Annex B of Itad and OPM (2015), 'Empowerment and Accountability Annual Technical Report 2015: Final Draft Version', Brighton, Itad, May
[3] Pawson et al, op cit

Figure 1: Macro evaluation components and methods



Reliability: Ensuring the repeatability/ replicability of findings

The first robustness principle of reliability ensures that the findings generated through the chosen research method are repeatable over time, across projects and across researchers.

Applying this principle to our realist evaluation method means ensuring that the QCA 'conditions' (comprising contextual factors, project mechanisms and project outcomes) are identified and scored (using QCA binary scoring) in a replicable manner and that the emerging patterns/ causal configurations are then interpreted through narrative analysis in a replicable manner by any researcher using the same method.

In practical terms this means establishing a clear and replicable tabulated coding and rubric system that can be systematically applied by a group of researchers with shared conceptual understandings of the conditions involved. This is what gives the coding its transparency and openness to external scrutiny and challenge.[4] These rubrics use a mix of proxy indicators and extracted qualitative data:

- The proxy indicators for project contexts are selected from nationally-comparable governance indexes and are used for standard binary measurements of the presence or absence of various contextual conditions (such as the strength of civil society or the openness of political society). These scores are reductionist but unambiguous, dividing the project set cases into two groups (1 or 0, with no case slipping between the two);
- Extracted qualitative data are used for additional binary coding: to code for the presence or absence of project mechanisms (such as support to local dialogue or capacity building of media) and to code for evidence of achievement of project outcomes (such as strengthened civil society or improved service delivery). The

[4] These raw data will be available for scrutiny by a peer review group established with DFID, and we are open to discussions about how much public access we will allow for wider scrutiny.

extracted qualitative data are included in the relevant tabulated cell, accompanied by a summary statement that justifies the binary score applied;

- We will also test the replicability of our findings through sensitivity analysis of our QCA results. We will randomly add and remove conditions and cases from our models, and change calibration thresholds. The ease and extent to which this changes our results will give us an indication of the sensitivity of our QCA results. We will identify what constitutes acceptable versus excess sensitivity and will make this clear when we report on the results of these tests.

In order to increase our confidence that we have applied replicable scorings to the conditions and that the QCA analysis will therefore generate replicable sub sets of projects with shared 'causal configurations' that can be subject to interpretive analysis of cause and effect using narrative analysis (see internal validity discussion below), we will subject the coding and tabulating process to triangulation. This involves as a first step *ex ante* work of normalisation amongst researchers through piloting and spot-checking. Once work begins on the main sample, the triangulation process involves random cross checking between researchers of the coding of conditions, including the extraction and summarising of relevant qualitative evidence.

Internal validity: Increasing the confidence that we can place in identified cause and effect relationships

Reliability alone is not sufficient for ensuring a robust research methodology. We may be very confident that we will get the same result if we repeat the measurement but it does not tell us whether and how a given intervention is contributing to changing outcomes. Internal validity shows that the researcher has evidence to suggest that an intervention had some effect on the observations and results.

Establishing internal validity within our combined methods approach will involve first being confident about the causal configurations established by QCA and second being confident about our deeper interpretation of those configurations using narrative analysis. Hence:

- We will ensure first that the QCA analysis of the coded conditions (described under 'Reliability' above) is followed using a standardised and transparent protocol that is open to general external scrutiny and to specific scrutiny through a peer review panel established with DFID for this study;
- we will further ensure that sample sub sets, identified to explore shared causal configurations, are established with clear criteria for their formation. In QCA terms these are causal configurations of 'necessary' and 'sufficient' conditions associated with given outcomes.[5]
- For each causal configuration we will ensure that the selection of cases for in-depth, interpretive (narrative) analysis is transparent. We will identify three clusters of cases to subject to in-depth analysis:
    1. Cases that exemplify the configuration of conditions associated with a given outcome of interest. ('True Positives');

[5] We will express our findings in terms of necessity, sufficiency or INUS relations – consistently with multiple-conjunctural causal inference models.

2. Cases that are inconsistent, having the same configuration of conditions but with outcome absent ('False Positives'[6]);
3. Cases that are inconsistent, having the same outcome but with the shared configuration of conditions absent ('False Negatives').

- Within each of these categories there may be too many cases to subject all of them to narrative analysis. We will therefore sample from these cases transparently for the following clusters of cases and will select a minimum of three cases per cluster[7]:
  1. True positive cases: In order to find any likely causal mechanisms connecting the conditions that make up the configuration we will look for 'modal cases', i.e. those that have maximum similarity with all other cases in this group. We will use the 'Hamming distance' method of establishing similarity to find this type of case.[8] Once a plausible causal mechanism is found, we will check to see if it can also be found in the most 'marginal' cases in this group i.e. those with least similarity with all others (identified again using the Hamming distance method);
  2. False positive cases (if present in the identified causal configuration): We will select modal cases using the same method. We would expect to find the same causal mechanism to be present in these false positive cases but to find some other factors that are blocking it from working delivering the outcome;
  3. False negative cases (if present in the identified causal configuration): We will select modal cases using the same method. Given the absence of the same configuration of conditions we would not expect the same casual mechanism to be present in these cases.
- It is important to flag here that we will be selective in our application of this method of within-case analysis. We will prioritise within-case analysis based on our recognition of: (a) resource limitations, (b) data limitations and (c) stakeholders' views of which configurations are high versus low priority for this kind of analysis.
- We will then subject these causal configurations to within-case analysis with the following objectives[9]:
  1. **Verification** that the attributes of a project are actually those that are ascribed to them in the data set used in the QCA analysis. Given the procedure described above for coding, few errors should be expected, but will be addressed if they occur;
  2. **Enlivening** the QCA coding through the construction of simple readable narrative which connects the conditions in the configuration in a way that is both plausible and respectful of the facts;
  3. **Excavation** to establish if there is a 'real life' causal mechanism or explanatory model that connects the events described by the configuration of conditions found via QCA.
- We will increase the trustworthiness of the causal inference in our narrative analysis through demonstrating the 'rigorous thinking'[10] in our narrative analysis. We will map

[6] Any causal mechanism identified within QCAs consistent cases (True Positive cases) should not be also present in the inconsistent cases (False Positive cases) (Rick Davies, pers. comm.).
[7] Assuming one dominant configuration per hypothesis.
[8] We will retain the option to prioritise cases with higher quality evaluative evidence for narrative analysis if these cases are also close to the modal case profile.
[9] Rick Davies (pers. Comm..).
[10] On the distinction between rigour as statistically verifiable attribution and rigour as 'quality of thought, see Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations.* (Working Paper No. 38), London, Department for International Development;

our inferences against the evaluative evidence available and consider the strength of our hypotheses against possible rival explanations. This approach represents a simplified adaptation of the empirical tests sometimes applied in the qualitative evaluative method of 'process tracing'[11]

- We will further strengthen our confidence in the verifiability of these emerging explanatory models by subjecting them to cross-checking and interrogation by at least one other researcher, who will review the evidence cited and its interpretation. This internal challenge function -- the basis of achieving trustworthiness in qualitative research[12] -- will enable us to increase our confidence in the internal validity of our interpretations.

External validity: generalising results beyond the immediate study population

The third and final principle that we apply to the macro evaluation research process is that of external validity. This increases our confidence that we can generalise our findings beyond the sample group and apply them to a larger population of interest.

In conventional empirical research external validity is established with a probability-based (random) sample that is sufficiently large to capture the variability of the 'population universe' (in this case the total Social Accountability project portfolio) under study.

The process of constructing project sets for the macro evaluation is described in the methodology annex (Annex B) of the E&A Annual Technical Report 2015.[13] This makes it clear that we have not been able to construct a probability-based sample from the Social Accountability project portfolio as we are limited to those projects whose evaluative content is quality-assured (as of summer 2014, 77 out of a total of 180, although this may increase slightly, with the addition of annual reviews and evaluation reports completed in the past year).[14] This in itself introduces an unavoidable bias towards those projects, which are well documented and evidenced. However, for the next round of analysis, we will include as many as possible of the 77 quality-assured projects to increase the coverage and breadth of our knowledge relating to the project portfolio. We have started the process of conducting a final data quality screening and are confident that the final number of quality-assured projects will be in the region of 50, and therefore within the budgetary ceiling of this analysis. This approach will increase our confidence that we have captured the variability of 'causal pathways' identified by QCA and explored through narrative analysis across the Social Accountability project portfolio. Moreover, since we are not sampling and using all projects with sufficient data quality, other sources of bias are relatively limited. Other possible biases may arise from geographically-prioritised or politically-driven selection of projects for additional evaluation or extra scrutiny by DFID.

To explore possible biases, we analysed the extent of the representativeness of this project set by mapping the project set profile onto the total project population using the portfolio

---

White, H., & Phillips, D. (2012). *Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework* (Working Paper No. 15), International Initiative for Impact Evaluation (3ie).

[11] Collier D (2011) "Understanding Process Tracing"*, PS: Political Science and Politics*, 44:4 pp 823 -830, University of California, Berkeley.
http://polisci.berkeley.edu/sites/default/files/people/u3827/Understanding%20Process%20Tracing.pdf

[12] Lincoln, Y. S. and E. G. Guba (1985). *Naturalistic Inquiry*, London, Sage

[13] Itad and OPM (2015), op cit

[14] The three quality assurance criteria of triangulation, transparency and contribution are described in Annex B.

synopsis descriptive data. We compared our project set of 77 quality-assured projects to the overall population of 180 Social Accountability projects on descriptive criteria such as geography, duration, budgets, etc. We also compared the distribution of DFID outcome scores where available, which provided us with a preliminary indicator of possible positive or negative bias. Our comparative analysis confirms that the sample is highly representative against these criteria. We will detail this comparative analysis in an annex of the next technical report.

When identifying and interpreting causal configurations of conditions that are associated with a specific outcome, we will focus on those conditions that are consistently displayed by a large number of cases. This will increase our confidence of interference and allow us to identify relatively generalisable findings[15]. To facilitate this, we will keep the ratio of conditions to cases small[16]. If findings are illustrated by a large number of cases with few inconsistencies, this will provide an indication of generalisability.

Finally, our realist synthesis approach will allow us to explain the *absence* of external validity in individual project causal mechanisms that we identify. We will be able to identify and interpret those projects – particularly through our case selection method of identifying false positive or false negative cases -- where causal mechanisms are too contextually specific to have external validity in order to share lessons on what mediating aspects of project context ensure that explanatory models are *not* generalizable to a wider population of projects.

[15] However, we will also analyse outlier configurations where they offer interesting learning opportunities.

[16] We will also look at some of the tables suggested by Marx and Dusa (2011), which intend to calculate probabilities of obtaining contradictory configurations for given numbers of cases and variables. However, we are aware of the limitations of this approach and will only use it where best applicable.