Centre for Development Innovation

Realist Evaluation: an overview

Report from an Expert Seminar with Dr. Gill Westhorp

Gill Westhorp
Ester Prins
Cecile Kusters
Mirte Hultink
Irene Guijt
Jan Brouwers

Seminar Report

Learning by Design

WAGENINGEN UR
For quality of life

Context, international cooperation

Wageningen UR Centre for Development Innovation (CDI) works on processes of innovation and change in the areas of secure and healthy food, adaptive agriculture, sustainable markets and ecosystem governance. It is an interdisciplinary and internationally focused unit of Wageningen University & Research centre within the Social Sciences Group.

Through facilitating innovation, brokering knowledge and supporting capacity development, our group of 60 staff help to link Wageningen UR's expertise to the global challenges of sustainable and equitable development. CDI works to inspire new forms of learning and collaboration between citizens, governments, businesses, NGOs and the scientific community.

More information: www.cdi.wur.nl

Innovation & Change

Ecosystem Governance

Adaptive Agriculture

Sustainable Markets

Secure & Healthy Food

Dr. Gill Westhorp has worked in the community services and health industries for over 25 years, in service delivery, policy, training, management and consultancy roles, in both Government and non-Government sectors. She has worked in community health; community development; education, employment and training; and crime prevention. She works for Community Matters, a consultancy business based in South Australia. For more information: www.communitymatters.com.au

Context, international cooperation (Utrecht, the Netherlands) is a development organisation understanding itself as a social enterprise. We are innovators who use business principles that aim to achieve fundamental social change and generate revenue. Currently Context has 14 staff members, in addition to a number of external Programme Associates. We position ourselves between the development and academic communities, and try to facilitate dialogue between policymakers, academics and development practitioners. Context staff members and Associates have a sound academic background as well as working experience in development practice. For more information: www.developmenttraining.org

Learning by Design is the company name under which Dr. Irene Guijt operates. She works independently on assessing results in social change, such as through the BetterEvaluation initiative that is being beta-tested and the Big Push Forward. She advises and researches on, and facilitates learning processes and systems in rural development and natural resource management, particularly where this involves collective action. Work over the past twenty years with multilateral organisations and many development NGOs has focused largely on strengthening critical reflective thinking to strengthen pro-poor development.

Realist Evaluation: an overview

Report from an Expert Seminar with Dr. Gill Westhorp

Gill Westhorp
Ester Prins
Cecile Kusters
Mirte Hultink
Irene Guijt
Jan Brouwers

Seminar Report

May 2011
Project code 8141100500
Wageningen UR Centre for Development Innovation

Realist Evaluation: an overview
Report from an Expert Seminar with Dr. Gill Westhorp

Edited by Dr. G. Westhorp
Prins, E.
Kusters, C.S.L.
Guijt, I.
Brouwers, J.H.A.M.

May 2011
Centre for Development Innovation, Wageningen University & Research centre; Context, international cooperation, Learning by Design

This report summarises the discussions and presentations of the Expert Seminar 'Realist Evaluation', which took place in Wageningen on March 29, 2011. The Expert Seminar was organised by the Wageningen UR Centre for Development Innovation in collaboration with Learning by Design and Context, international cooperation.

Cover photos
Guy Ackermans

Orders
+ 31 (0) 317 486800
info.cdi@wur.nl

Table of contents

Executive summary iv

1 What is Realist Evaluation? 1

2 Realist Evaluation in the broader context of theory-based evaluation 2

3 Assumptions about reality and their meaning for evaluation 3

4 Assumptions about reality and their meaning for evaluation 4

5 Context 8

Appendix 1 – More information 13

Executive summary

This report summarises the discussions and presentations of the Expert Seminar 'Realist Evaluation', which took place in Wageningen on March 29, 2011. The Expert Seminar was organised by the Wageningen UR Centre for Development Innovation in collaboration with Learning by Design and Context, international cooperation.

The report describes what is Realist Evaluation (RE), which mainly focuses on the question: *'what works for whom, in what contexts, in what respects and how'.* It also places realist Evaluation in the broader context of evaluation, where *Evaluation theory* deals with how things can be evaluated and what can be known about the results.

The report also discusses assumptions about reality and their meaning for evaluation. Evaluation is not simple and outcomes are not linear. There are always multiple and competing mechanisms operating. Furthermore, assumptions about knowledge and its meaning for evaluation are discussed, and how positivism, realism and constructivism have different implications for evaluation due to philosophical differences.

The report concludes by looking at context and how RE and action research can be complementary, but also how program logic and RE can be complementary. Also methodology for RE and RE in complicated/complex programs and systems are discussed. It concludes with describing when and when not to use RE.

1 What is Realist Evaluation?

The term 'realist evaluation' is drawn from Pawson and Tilley's seminal work, Realistic Evaluation (1997). It is an approach grounded in realism, a school of philosophy which asserts that both the material and the social worlds are 'real' and can have real effects; and that it is possible to work towards a closer understanding of what causes change. The word 'Realistic' in the book title is a play on words: "real" because it refers to work done in the real world; realist because it is grounded in realism; and "realistic" because it is a form of applied research, which is "pursued to inform the thinking of policy makers, practitioners, program participants and the public".

Realist Evaluation changes the basic evaluation question. It does not ask what works, or does this work. It asks *'what works for whom, in what contexts, in what respects and how'.* A realist approach assumes that programmes are 'theories incarnate'. That is, whenever a program is implemented, it is testing a theory about what 'might cause change', even though that theory may not be explicit. One of the tasks of realist evaluation is therefore to make the theories within a programme explicit, by developing clear hypotheses about how, and for whom, programmes might 'work'. The implementation of the programme, and the evaluation of it, then tests those hypotheses. This means collecting data, not just about programme impacts, or the process of programme implementation, but also about the specific aspects of programme *context* that might impact on programme and about the specific *mechanisms* that might be creating change.

Example 1: An apple a day keeps the doctor away (Prof. Patricia Rogers, with permission)

The proverb 'an apple a day keeps the doctor away' can illustrate Realist Evaluation. A program logic model would see 'apples' and 'people in poor health' as inputs and apples being eaten as an immediate result (output). The short term outcome would be improved nutritional status and the longer term outcome would be improved health status.

Realist Evaluation tries to investigate how the occurred change happened. In this case, did change occur because apples are red? Because apples contain vitamin C? Or because they are healthier than junk food? To test these different theories, it is important to distinguish the active ingredient that caused the change (for example vitamin C in case of the apples), quercetin (an anti-oxidant found in naturally-occurring red food colours) or substitution (in case of reduced junk food).

Realist Evaluation also asks for whom things work. Vitamin C will only work for those who have vitamin C deficiency and will cause a reduced incidence of scurvy. Increased levels of quercetin will result in reduced incidence of cancer, heart diseases and, for men, reduced inflammation of the prostate. Decreased consumption of junk food works for overweight people and will result in reductions of obesity-related health conditions. For whom it works, and what the exact outcome will be, depends on *how* it works. Outcomes therefore depend on both the mechanism and the context.

It is important to note that the 'active ingredient' (an analogy for 'program mechanism') does not mean 'an aspect of the program activity' (here, apples). If the change was due to vitamin C, one could cause the same effect by using oranges. In case of the red color, one could use red onions, or if the change was related to not eating junk food, one could use carrot sticks. Realist Evaluation is about finding out how things work. Once that is understood, a program can be tailored for a specific context: – a person working in an area with no apples knows whether to choose carrots or red onions to achieve their goals.

2 Realist Evaluation in the broader context of theory-based evaluation

What does the word 'theory' mean in theory-based evaluation? Four kinds of theory can be distinguished: philosophical theory, evaluation theory, programme theory and substantive theory. The deepest level, philosophy, deals with assumptions about the nature of the world and how we understand it. These can never be proven. *Philosophy* includes theories about 'what exists' (ontology), and what can be known about it (epistemology). These theories or beliefs influence all the other kinds of theory. *Evaluation theory* deals with how things can be evaluated and what can be known about the results. For example, some people argue that different philosophies underpin the quantitative and qualitative paradigms in evaluation. *Programme theory* describes the theory built into every programme. There are different ways of representing program theory, including programme logic and theory of change. *Substantive theory* operates within a particular domain or discipline, for example psychology.

Realist Evaluation can be represented at all four levels. On the level of philosophy, Realist Evaluation is grounded in realism and in systems theory (for human services programs, a particular kind of systems theory, called complexity theory, is often useful.) Realist evaluation itself is a particular evaluation theory. Realist Evaluation develops a particular kind of programme theory, structured as Context- Mechanism-Outcome Configuration (CMOCs). Lastly, substantive theory feeds realist evaluation with clues on the mechanisms (e.g. the combination of reasoning and resources that enable a programme to 'work') through which programs work and the contexts in which they will work.

3 Assumptions about reality and their meaning for evaluation

A fundamental belief in realist ontology is the existence of a 'real' world, which exists independent of (or interdependent with) our interpretations of it. (It is independent of our interpretations of it, but how we interpret it influences our actions, which in turn can influence reality: this is why some might say 'interdependent with' our interpretations.) In this ontology, the material and social world are both real: things are considered 'real' if they can have a real effect in the world. Another assumption is that everything in this world is organised in systems, which in turn are embedded in larger systems and connected to other levels. For example, in the material world, cells can be seen as being part of sub-systems that make up the larger system of the human body. In the social world, on the other hand, a human being forms part of subsystems such as a family and community. Similarly, ideas are built in to larger systems of belief, which in turn form part of culture. Social systems are open: elements can move in and out of the system. For example, family members can leave, others can marry into the family. Everything is embedded into other levels and all the systems interact with each other. As a result, any event has many causes, and at the same time may have many consequences. This also means that every outcome of a programme is a result of multiple causes; and that every program may have many different outcomes.

With regard to evaluation, these assumptions about reality mean that programs and policies are real and can have real effects - positive, negative, intended and unintended. Social programs operate as open systems in which all levels are interacting. Programmes change systems and systems change programs. This means that evaluation is not simple and outcomes are not linear. There are always multiple and competing mechanisms operating. Mechanisms also interact with their context, which is why a programme can generate 'x' outcomes in one setting and 'y' outcomes in another.

4 Assumptions about knowledge and its meaning for evaluation

Experiences and understandings are filtered through the human brain and language. Every human brain is different and interprets differently. Language also shapes how a person interprets his experience. Some people argue, therefore, that it is impossible to state that one person is right and another wrong. However, realists believe that the real world 'regulates' and sets boundaries on what is *reasonable* to believe, and that knowledge can be improved over time.

With regard to evaluation, this means that there is no absolute certainty about evidence to date, and also no absolute certainty about evaluation findings. Evidence is just 'the best we think we know so far'. Evaluations should therefore be iterative and knowledge should be treated as cumulative, across policies and programmes. Realist evaluation is designed to enable improving understanding over time: how can learning from one programme be taken to another?

Philosophical assumptions about causation and its meaning for evaluation

In realism, the term "mechanism" relates to causation. Powers or processes which generate events, or patterns of events, can be seen as a 'causal mechanism'. Mechanisms operate at all levels of reality and the outcomes of any mechanism are usually at a different level from the mechanism itself. Mechanisms, therefore, cannot usually be directly observed; they need to be hypothesized and tested. Whether mechanisms 'fire' or not depends on the context.

Example 2: The Tennis ball - Realism in a nutshell

A person standing on planet Earth has a tennis ball in his hand. When he opens his hand, the ball will fall, due to gravity. In this case, gravity is the mechanism, the opening of the hand is an analogy for the programme strategy. When the same person is placed into space and opens his hand, the tennis ball will stay in the same place, as gravity is too weak to move the ball. In a third situation, the person is back on earth, but under water. When opening his hand, the ball will float. Gravity still exists, but the ball is also subject to buoyancy. Multiple mechanisms (gravity and buoyancy) are operating, but buoyancy is stronger than gravity when under water.

In all three cases, the programme strategy (opening of the hand) was the same, but the outcomes were different. Different mechanisms fired or outweighed each other. Realist Evaluation therefore tries to find out which mechanisms are present, which ones fired and in what circumstances (contexts) they work.

With regard to evaluation, one of the tasks of an evaluator relates to the attribution of outcomes. Is this particular program responsible for the outcomes we see? This means that evaluators are concerned with causation. A Realist Evaluation tries to identify the mechanism that 'fired' and to understand what caused (or, given multiple causation, at least contributed to) the outcomes.

Philosophical differences: Positivism, Realism and Constructivism

The philosophical differences between Positivism, Realism and Constructivism can be summarised as follows:

	Positivism	**Realism**	**Constructivism**
Ontology	There is an objective reality, which exists independent of us.	Material & social reality – we interact with reality.	Subjective reality – we 'create' reality
Epistemology	Truth and final knowledge exists.	No final truth or knowledge, but improvement in knowledge is possible.	No way to choose between interpretations. What we jointly believe is true.
Causation	Constant conjunction, linear causation. Programs cause outcomes.	Mechanisms operating differently in different contexts generate patterns of outcomes.	Co-constructed interpretations lead to actions and outcomes.
Implications for evaluation	Evaluators 'tell facts'. Context factors should be eliminated: Randomised Control Trials/ Quasi-experimental methods.	Evaluators explain how and where programs generate outcomes. Mixed methods (qualitative and/or quantitative).	Evaluators describe stakeholder interpretations. Qualitative methods

Programs and mechanisms

Programmes are intended to create change. In order to create change, programmes make use of an (implicit) theory about how change might occur. The task of an evaluator is to test and understand the program theory. Realist program theories provide a description of mechanisms that are expected to operate, how context will affect whether or not mechanisms operate, and what outcomes are expected.

The traditional view is that programmes are the active ingredient that causes change. Realist evaluation turns this around. It says that programs offer resources or opportunities, but it is the decision-making of participants that causes the outcomes. (It is still necessary for the program to be implemented properly for the program to be effective.) Program 'mechanisms' relates to 'how' programmes work. The term 'mechanism' refers to how programs change people's decision-making: what people do in response to the resources that the program provides.

Example 3: Closed circuit television camera (Pawson and Tilley, 1997)

In the book 'Realistic Evaluation', Pawson and Tilley provide the example of using closed circuit television (CCTV) to decrease theft from cars in car parks. The camera itself does not actually prevent theft. It might influence the 'reasoning' of potential offenders, of people who park in the car park, or of security guards or police. For example, the offender may reason that the risks of being caught are too high and decide not to offend (a deterrence mechanism). Alternatively, they could decide to offend in a different car park (a displacement mechanism). The CCTV might also attract cautious car owners to park in that car park, who remove their valuables and lock their cars, making theft more difficult (a target hardening mechanism). In the book, Pawson and Tilley describe eight possible mechanisms and six different contexts that might affect whether and how the cameras 'work' to reduce theft.

In summary: programs provide resources or opportunities, to which people respond in different ways, generating patterns of outcomes. Reasoning + Resources = Mechanism.

Choice making never happens in vacuum and is always enabled and constrained at two levels. At micro (individual) level, choices are affected by people's beliefs, resources, expectations, experiences, attitudes, resource availability and so on. On macro (society) level, decisions are shaped by social environments, culture, norms and other social forces.

Example 4: Patterns in social behaviour – Strangers in a lift

Why do strangers in a lift always face the front? One theory could be that it might seem rude or bad mannered to stare at another. Another theory could be that the behaviour relates to protection: making eye contact could encourage people to communicate, and one does not know whether the other person is trustworthy. This theory, however, might apply more to some groups than for others. Men, for example, might feel less intimidated then women. These patterns of behaviour do not only apply in lifts: they can apply wherever strangers are in close proximity. They provide an example of the ways social norms shape behaviour, differently for different groups, within and across social settings.

Different kinds of programs trigger mechanisms at different levels of reality. Consider the example of changing an individual's behaviour. Cognitive Behaviour Therapy (CBT) operates at a deep level, changing how the limbic and cortical areas of the brain are structured and how they work. A second type of programme - education or training - also changes the brain, but through the mechanism of raising awareness, skills, knowledge, or building confidence. Thirdly, a road construction scheme may improve economic capital (for example access to markets) and thus enable new behaviour. And finally, a community development program may enable individual change by creating social capital, through the mechanism of changed norms or beliefs.

Example 5: Home visits by nurses (David Olds et al)

In a programme which aimed to improve health and life outcomes for babies and their mothers, nurses were contracted to visit families with newborns and provide them with support. The program was evaluated and demonstrated to be successful, but some people thought it might be more effective if 'peers' (community members) provided support instead of nurses. A later programme compared outcomes for two groups of families: one supported by nurses and the other supported by community members, who received twice the training and supervision of the nurses. Outcomes for families visited by nurses were better than those of families visited by community members. It has been theorised that two mechanisms account for these better outcomes. One is that nurses are credible sources of information about new babies and parents are more likely to follow their advice. The second is lack of stigma. In the state where the trial was conducted, all new mothers are visited by nurses, but not all mothers are visited by community members (suggesting that those who are visited must be doing something wrong). According to this theory, "high credibility, low stigma" provision of information and support is what needs to be replicated – and in other communities, depending on local context and culture, nurses may or may not be the best way to achieve that.

Program activities and mechanisms are not the same thing. Different programs, triggering different mechanisms, might lead to the same outcome. However, the same program might trigger different mechanisms in different contexts, or for different groups of participants, and therefore create different outcomes.

Example 6: Programmes, mechanisms and outcomes (examples from various articles by Pawson and Tilley)

Breakfast clubs

Breakfast clubs were originally expected to work by improving nutrition for undernourished children, thus enabling them to perform better in school. However, breakfast clubs might also work by allowing children with high energy levels to let off energy before starting school. Or they might work by providing safety and affection for children in difficult home situations, enabling them to calm down before school starts. A programme, therefore, may work through many different mechanisms, which might not always be those intended.

Mandatory arrest for domestic violence

Mandatory arrest caused domestic violence rates to rise in some disadvantaged communities and to fall in other (less disadvantaged) communities. That is, the same programme generated different outcomes in different kinds of communities. It has been theorised that in the more advantaged communities, arrest caused a feeling of shame within the offenders, who sought help to address their violence, but in more disadvantaged communities, arrest provoked anger, which led to increased domestic violence.

Naming, shaming, faming

A number of programs provide public information about products or services, expecting that this will enable consumers to make informed decisions, which will in turn prompt producers or providers to increase the quality of the product or service. For example, publishing rankings of the safety of different makes of cars prompted some consumers to purchase safer cars. As a consequence, manufacturers of unsafe cars decided to re-engineer their cars. Through naming (shaming and faming), the market operated differently to produce safer cars and a *social good* was generated.

If the same strategy is used with schools (making lists of best and least performing schools), some (better off) parents might decide to put their children in better schools, leaving a greater concentration of poorer families in poorer schools. A greater concentration of disadvantage is enough on its own to cause a decrease in educational outcomes, so the program can generate greater inequality in education outcomes, with negative outcomes for the most disadvantaged. Here, a *social ill* is generated. The 'policy context' and the 'market context' affect how the mechanism works.

5 Context

In Realist Evaluation, interactions between context and program mechanisms determine the outcome. Context refers to features of participants, organisation, staffing, history, culture, beliefs, etc. that are required to 'fire' the mechanism (or which prevent intended mechanisms from firing). Population groups ('for whom' a program works) are one aspect of the context. Other contextual elements might include geographic and community setting, nation, culture, religion, politics, historical period, events, organisational setting, key attributes of workers, and so on.

Example 7: Potential elements of context – A handful of examples

Geographical setting
In Australia, sniffing petrol in order to get high is a recurrent problem with some members of some Aboriginal communities. Many programmes were developed to reduce the number of petrol sniffers, using mechanisms such as building pride and cultural identity. One program that worked well was conducted a long distance from any available petrol, suggesting that the program mechanisms were best able to fire in a particular geographic context (that is, a long way from petrol).

Community settings
In order to discourage stealing of property, some programmes focused on engraving names etc. into property. This worked best in relatively small communities, where marketing rates of property were high, publicity about the program was high, and relatively few offenders were present. In large metropolitan cities, the programme worked less effectively.

Historical periods
The point in time at which a program is introduced can affect its effectiveness. In crime prevention, as new strategies are developed to prevent crimes, offenders develop new ways to beat those strategies. Older strategies may not be effective in new situations.

Events
As a result of 9/11, a Muslim support service almost collapsed as racism towards Muslims increased and Muslims were afraid to be seen accessing their services.

Organisational setting
Rape is a disempowering experience. Rape counselling therefore generally focuses on re-empowering the victim. Counsellors reported that a rape programme in prison did not work well because the environment in which the intervention took place was disempowering.

Population groups
Gender roles can influence the population or subpopulation for the program. For example, in some agricultural research for development programmes, women would only participate in aspects of the research that related to their usual roles in farming.

In order to assess which factors in the context affect how programs work, the context has to be addressed in the evaluation design. It is important to formulate questions for identifying the context. Questions might be: For whom do you think this will work best? And for whom will the programme not work? When thinking about the times the programme worked best and when it did not, what characterised

the differences? Or the list of contextual features above can be reviewed, asking 'what matters about the context, for the way we expect this to work'?

With regard to the selection of key population groups, it is important to question why these subgroups would matter, based on theory (not just demographics). One has to be able to suggest why a program will work differently for the different population groups. One should then determine indicators to distinguish between subgroups, collect data about the sub-groups, and analyse programs outcomes by subgroup. The same should apply to key features of context that are hypothesized to affect whether and how a program works. One should specify why those features matter (based on theory), determine indicators for those features, collect data and analyse outcomes for the different contexts.

Example 8: Participatory agricultural research: An 'innovations' mechanism

The underlying theory for some participatory agricultural research for development programs is that a collaborative relationship between farmer and researcher will support innovation. The relationship allows the researcher to understand the local context and the farmer to understand the research process. A joint understanding of the problem is developed. Combining the farmer's and researcher's knowledge generates new ideas, developing solutions fitting the local context. Those potential solutions are then tested by local farmers. At the same time, involving the farmers increases farmers' belief in the effectiveness of the product and ensures that they know how to use it. Application of the solution leads to (for example) improved yields and therefore to improved livelihoods.

One way to identify important aspects of context is to consider 'threats to the theory'. In this example, innovation theory suggests that innovation requires a constructive relationship between the stakeholder groups who bring different knowledge to the process (here, the researcher and the farmer). But what might happen if the relationship fails? Researchers do not always have collaboration skills; or marginalised people with key knowledge might be excluded from the programme. Furthermore, donors or research institutions do not always actively support and provide resources for the additional costs of participatory research. Another threat might be that farmer participants might turn out to be better off farmers, who benefit more and faster from research than others, which could result in increased inequality.

Once potential threats to the program theory have been identified, data can be collected about these critical aspects, to assess whether or not they affect how the program works in different contexts.

Realist Action Research and Realist Program Logic

Realist evaluation is a way of thinking (a kind of 'logic of evaluation') rather than a model of evaluation. The core ideas need to be included in an evaluation design. They can also be incorporated within other existing evaluation approaches. Two examples of integrating realist ideas into existing approaches are 'realist program logic' and 'realist action research'.

Combining action research and Realist Evaluation

Action Research and Realist Evaluation can be combined. The action research cycle remains unchanged, but the nature of the questions that are asked at each stage of the process is different. The questions are amended to take account of context and mechanism, the theory that is tested is structured as a realist theory (CMOC), and data is collected to test each aspect of that theory. The following matrix shows the added value of Realist Evaluation (RE) for Action Research (AR) and vice versa.

RE adds to AR	AR adds to RE
A structure for program theories (CMOC's)	Overt cyclical/responsive structure
Focus on *how* and *for whom* programs work	Overt participatory approach ('the researcher's theory' – 'the program stakeholder theory')
Guidance about 'what to look for' about context (what affects whether mechanisms fire)	Acceptability to (some) services
A structure for portable lessons from AR	

Combining Programme Logic and Realist Evaluation

Programme Logic (e.g. Logframe) and Realist Evaluation can be combined. The first step is to map out the basic program logic. Then the 'realist components' (mechanisms and contexts) can be added. Realist Evaluation can help us to understand what activities trigger which mechanisms, and how they lead to certain outcomes in certain contexts. The following matrix shows the added value of Realist Evaluation (RE) for Programme Logic (PL) and vice versa.

RE adds to PL	PL adds to RE
Explicit focus on causal mechanisms	PL provides access to or how to use assumptions other than context and mechanism
Identification of short-term outcomes as mechanisms	Planning and management tools (e.g. program resources, timeframe)
Additions to the structure for program theories	Explicit focus on program processes (distinguishing program failure from theory failure and identifying additional mechanisms)
Guidance on 'what to look for' about context	Structure for monitoring progress
A structure for portable lessons from PL	

Methodology of Realist Evaluation

As a starting point for Realist Evaluation, the evaluation questions should be realist in nature. Once the programme theory is understood, tentative Context-Mechanism-Outcome Configurations (CMOCs) can be developed. Realist theory suggests that different stakeholders have different information because of their different roles in the program. Different information can be collected from different stakeholders (such as program designers, participants, service providers, managers, policy staff, and researchers) to develop these hypotheses.

As a next step, data items should be identified for the most important element of the hypotheses and data collection methods should be determined. The methods used in Realist Evaluation can be both quantitative and qualitative in nature. Quantitative methods are often useful with regard to context. For example, it is easier to compare across population subgroups using quantitative data. Quantitative data can also be used to test for some kinds of outcomes, and for some sorts of mechanisms, assuming that are clearly

hypothesised (e.g. how the manifests in reality). Qualitative methods are useful for exploring and developing hypotheses, investigating mechanisms (especially when these are not well understood), and to identify unanticipated elements of the context and unanticipated outcomes.

Once the data have been collected, they need to be analysed. The logic of analysis in realist evaluation is "intra-program, inter-group (or inter-context)" comparison. That is, realist program theory expects that there will be different outcome patterns for different groups or contexts within the program, and the analysis tests those theories. Control or comparison groups outside of the program are not required.

Realist Evaluation in complicated/complex programs and systems

A "complicated" program is a program that comprises of many (interconnected) parts. "Complexity" refers to the principles of complexity theory, including the idea that "simple rules guide interactions at the local level and generate complex patterns of outcomes at higher levels of the system". Some people have questioned whether realist evaluation can be used with complicated or complex programs. Strategies that can be used to apply a realist approach with complicated and complex programmes include "analysing a small slice of a complex pie" (this is Ray Pawson's suggestion. He also suggests considering the bit that is investigated or understood the least); 'layering' systems, mechanisms and/or theories to reflect the different levels at which a program is supposed to work; changing the level of abstraction (e.g. investigating how the various elements of interventions are supposed to work together rather than looking at each one separately); or looking for interactions and understanding the 'simple rules' that guide them.

Example 9: Changing student learning outcomes through changing teacher's practice

A programme intended to improve student learning outcomes by changing teachers' practice used four basic strategies: professional development, communities of practice, teacher mentoring and leadership engagement and support. These strategies were intended to change teachers' attitudes, confidence and skills in both mathematics and inclusive pedagogy, which would lead to more inclusive teaching behaviours, which would improve student engagement (especially by those who were more 'at risk' or excluded). As a result, student learning outcomes would improve and the 'achievement gap' between students would decrease.

"Teachers' attitudes, confidence and skills" needed to be defined. Different stakeholders had different beliefs about how the program would work (ie different mechanisms). Does increasing teachers' mathematical knowledge lead to more confidence to experiment with more inclusive pedagogy? Or do teachers change their attitude to their role and responsibility for learning outcomes, especially for excluded students? Or does increased skill in inclusive pedagogies lead to improved mathematic outcomes?

Once the theories had been identified, two of the three could be tested with a single, multiple choice question (using the stem "Which of the following statements is closest to true for you?"). It identified that teacher confidence in mathematics knowledge did not change, but that confidence in inclusive pedagogy did change. By using other data sets in the evaluation to triangulate and refine this finding, a statement could be made outlining the main mechanisms leading to changes in teacher practice.

When and when not to use Realist Evaluation

There are various reasons for undertaking Realist Evaluation. It is appropriate when the goal of the evaluation is learning about the program; when a policy idea is being introduced from another domain; when a program is being extended to a new population and/or when there are confusing patterns of findings from previous evaluations. Realist Evaluation is also appropriate when there have not been previous evaluations of a program or when there is a confusing pattern of outcomes within a programme.

Finally, if a programme has been found to be effective and might be replicated into another setting, realist evaluation can be used to identify what makes it work and therefore 'what needs to be replicated', or how it might need to be adapted for new contexts.

Realist Evaluation should not be undertaken when there is no real interest in learning; when resources (such as expertise or funding) are not sufficiently available; or when the purpose of the evaluation does not warrant it (e. g. cost benefit analysis at whole-of-program level or 'pure' process accountability). Finally, when the programme has not yet shown even early-stage outcomes, realist evaluation is not appropriate (the 'context-mechanism-outcome' approach requires outcomes data).

One criticism of realist evaluation is that it generates complex findings, whereas some stakeholders want 'simple answers'. However it is possible to provide different kinds of information from a realist evaluation for different stakeholders. Politicians and donors involved in funding programmes might need to know that different programmes work for different populations and that a mix of programs will be required. Policy staff in charge of designing and administering programmes might need to know which programs are required for which populations, and to understand broad Context-Mechanism-Outcome patterns for different programme types. System gatekeepers might need to know about the context and mechanisms of local programmes and which instruments to use to assess people for referral to appropriate programs. Service providers are the ones who have to tailor programs at the local level, and they need the most detailed knowledge about how and for whom programs work, and the skills to adapt them to local needs.

Appendix 1 – More information

Essential reading

Pawson, R. and Tilley, N. (2004) *Realist Evaluation*. Sage Publications Ltd, London (a chapter introduction rather than a book; see also http://www.communitymatters.com.au/gpage1.html).

Pawson, R. (2006) *Evidence Based Policy. A Realist Perspective*. Sage Publications Ltd, London (the seminal text for those interested in realist synthesis, ie realist literature review).

Tilley, N. (2000) *Realistic Evaluation, An Overview*.

Astbury, B. and Leeuw, F. (2010) *Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation*, American Journal of Evaluation, September 2010.

Also useful

Byng, R., Norman, I. and Redfern, S. (2005). *Using Realistic Evaluation to Evaluate a Practice-level Intervention to Improve Primary Healthcare for Patients with Long-term Mental Illness*. In: Evaluation, Vol. 11, pp. 69-93.

Henry, G.T., Julnes, G., and Mark, M.M. (Eds.) (1998). *Realist Evaluation: An Emerging Theory in Support of Practice.* Jossey-Bass Publishers.

Westhorp, G. (2011) *A brief introduction to realist evaluation.* http://www.communitymatters.com.au/gpage1.html.

Institute of Tropical Medicine, http://www.itg.be/itg/generalsite/Default.aspx?WPID=703&L=E.

This report summarises the discussions and presentations of the Expert Seminar 'Realist Evaluation', which took place in Wageningen on March 29, 2011. The Expert Seminar was organised by the Wageningen UR Centre for Development Innovation in collaboration with Learning by Design and Context, international cooperation.

Centre for Development Innovation
Wageningen UR
P.O. Box 88
6700 AB Wageningen
The Netherlands