WORKING PAPER / 2009.05

What role for qualitative methods in randomized experiments?

Martin **Prowse** and Laura **Camfield**

University of Antwerp

IOB
INSTITUTE OF DEVELOPMENT POLICY AND MANAGEMENT

Working Papers are published under the responsibility of
the IOB Thematic Groups, without external review process.
This paper has been vetted by Danny Cassimon, convenor
of the Thematic Group Impact of Globalisation.

Comments on this Working Paper are invited.
Please contact the author at **<martin.prowse@ua.ac.be>.**

I**nstitute of Development Policy and Management**
University of Antwerp

Postal address:	Visiting address:
Prinsstraat 13	Lange Sint Annastraat 7
B-2000 Antwerpen	B-2000 Antwerpen
Belgium	Belgium

tel: +32 (0)3 265 57 70
fax +32 (0)3 265 57 71
e-mail: dev@ua.ac.be
www.ua.ac.be/dev

WORKING PAPER / 2009.05

What role for qualitative methods in randomized experiments?

Martin **Prowse** and Laura **Camfield***

October 2009

* Martin Prowse, Institute of Development Policy and Management (IoB), University of Antwerp, Belgium, and Laura Camfield, Department of International Development, University of Oxford, UK.

The authors would like to thank Jos Vaessen for his comments, and Peter Davis for discussing his experience of using life history interviews. Needless to say, responsibility for any errors in fact or interpretation remain solely with the authors.

IOB
INSTITUTE OF DEVELOPMENT POLICY AND MANAGEMENT

University of Antwerp

TABLE OF CONTENTS

TABLE OF TABLES

ABSTRACT

The vibrant debate on randomized experiments within international development has been slow to accept a role for qualitative methods within research designs. Whilst there are examples of how 'field visits' or descriptive analyses of context can play a complementary, but secondary, role to quantitative methods, little attention has been paid to the possibility of randomized experiments that allow a *primary* role to qualitative methods. This paper assesses whether a range of qualitative methods compromise the internal and external validity criteria of randomized experiments. It suggests that life history interviews have advantages over other qualitative methods, and offers one alternative to the conventional survey tool.

1. INTRODUCTION

The effectiveness of development assistance has come under close scrutiny in recent years with the terms of debate shifting from the quantity of aid (Sachs, 2005; Easterly, 2005) towards improving quality. For example, the Paris Declaration of 2005 established five key principles to improve the effectiveness of aid flows, four of which focus on process issues such as increasing accountability.[1] A different approach to improving aid quality has been to improve the evaluation of aid's impact. Evaluation practices in many donor agencies have tended to focus on policy and strategy within country programmes, thus focusing on internal institutional issues as opposed to aid's impact on the well-being of the poorest (Foresti et al, 2007). Such an approach to evaluation contributed to a lack of evidence and consensus around that simplest of questions: what works? (Banerjee et al, 2007; Savedoff et al, 2006). In this respect, current evaluation procedures contributed to an attribution gap, whereby it has been difficult for agencies to assign improvements in well-being to specific policy interventions (see White, 2007b).

Such currents have contributed to an upsurge in interest in impact assessment and impact evaluation methodologies. Such approaches vary in terms of scale, the theoretical and conceptual frameworks employed, methodologies and research designs (and, often implicitly, epistemological and ontological beliefs), choice of measurement tool, choice of impact indicators, type of analysis, and value framework.

It is not our purpose to compare and contrast different approaches to evaluation, and certainly not to try and posit a hierarchy of techniques (which, in any case, must surely depend, *inter alia*, on the research questions in hand, resource envelope, and expertise of investigators). Instead, we focus solely on one type of impact evaluation – namely, randomized control trials – to assess the following question: what types of qualitative research method could play a primary role in an experimental design?

There are good reasons to attempt to answer this question. First, a number of relatively early contributions to the debate on impact evaluation barely mention qualitative methods. For example, the World Bank's Independent Evaluation Group report (2007) includes a nominal paragraph, as does a recent Asian Development Bank report (2006). The Evaluation Gap Working Group's influential report – Will We Ever Learn – offers even less (Savedoff et al., 2006). This is not particularly surprising. Randomized experiments have entered mainstream development debates through finding fertile ground in micro development economics, and it is quite unusual to find development economists who stray too far from their conventional survey measurement tool.

[1] These five principles are, first, to strengthen country ownership so that developing countries' governments set the agenda. Second, to increase donor alignment with government policies and management systems. Third, to increase donor harmonisation through improving co-operation and division of labour. Fourth, to focus on development results through better evaluation and learning. And fifth, mutual accountability, so that both recipient countries and donors are equally accountable for development results. Progress in implementing these important principles, and ensuring greater civil society involvement, was assessed in Accra, Ghana, in September 2008.

Second, the influential academics associated with the Poverty Action Lab at Massachusetts Institute of Technology (MIT), who are the leading proponents of randomized experiments, have, until recently, paid little attention to the use of non quantitative methods. But this is not to say they have been ignored. For example, in their study on the impact of reserving village leader roles for women on the provision of public goods in West Bengal and Rajasthan, Duflo and Chattopadhyay (2004) utilized participatory resource mapping (with ten to twenty villagers) and semi-structured interviews to ascertain village-level infrastructure investments and repairs. More recently, Karlan (2009) has rightly stated that "the decision about what to measure and how to measure it, i.e., through qualitative or participatory methods versus quantitative survey or administrative data methods, is independent of the decision about whether to conduct a randomized trial", and outlines further studies that utilize non quantitative methods (including Olken, 2007, and Karlan and Zinman, 2009). However, whilst the acknowledgment that qualitative methods can be utilised within a randomized experiment is to be welcomed (see Prowse, 2007, for an early discussion of this issue), Karlan (2009) says little about the precise qualitative or participatory tools that can be utilised.

And third, researchers associated with Network of Networks for Impact Evaluation (NONIE) and the International Initiative for Impact Evaluation (3IE) have long argued that qualitative methods or contextual research such as 'field visits' or descriptive analyses of the political and economic environment should play a complementary, if secondary role to rigorous quantitative methods.[2] Until now, though, researchers such as White (2008) and Leeuw and Vaessen (2009) have paid little attention to the possibility of randomized experiments that allow a primary role to qualitative methods (even though this is common within the related field of social policy, reviewed in Molloy et al. 2002, see also Gibson and Duncan, 2000; London et al, 2005; Lewis, 2007).

Therefore, and following Woolcock (2009), this paper assesses the extent to which different qualitative research methods could be used as the primary measurement tool within a randomized design. It does so in two parts: first, through assessing to what extent a range of qualitative methods can adhere to the basic characteristics of randomized design; and second, through assessing whether a range of qualitative methods compromise the internal and external validity criteria of a randomized experiment. The paper argues that two qualitative methods – life history interviews and semi-structured interviews – appear suitable, with the former holding particular promise.

The paper consists of five sections. The first introduces randomized experiments. The second outlines threats to their internal and external validity. The third looks at how qualitative research methods have been integrated within an experimental research design. The fourth section assesses a range of qualitative research methods in terms of whether they compromise the internal or external validity criteria of randomized experiments, and suggests there may be particular value in utilizing life histories as the primary measurement tool. The fifth section outlines future research avenues and concludes.

[2] Examples of this include the mixed-methods evaluations of the Bangladesh Integrated Nutrition Project and Kenyan agricultural extension services described in White 2006, p. 32-37.

2. WHAT ARE RANDOMIZED EXPERIMENTS?

Randomized experiments are designed and structured to answer a counterfactual question: how would participants' welfare have altered if the intervention had not taken place? They utilize a robust 'control' group who are not directly exposed to the intervention, and whose outcomes would probably have been similar to participants if the intervention had not taken place. This allows researchers to estimate the mean effect of a particular intervention across the 'treatment' group (indeed, both the group assigned to receive the treatment, and the group that received it – see Ravallion, 2009a). In this respect, we can offer a tripartite definition of Randomized experiments: first, they focus on the impact of an intervention on welfare/well-being outcomes of participants; second, they use counterfactual analysis; and third, they necessitate substantial primary research (in contrast to *ex post* counterfactual methods which often utilize pre-existing datasets to construct a control group).

Randomization overcomes important limitations in many non-experimental studies, in particular selection bias**.** In other words, that participants in any program are unlikely to be a random sample of the population as a whole (as programs are often 'targeted' at specific groups). Evaluating the efficacy of the program then becomes difficult, not least because a comparison group which *is* a random sample will not be comparable (White, 2007b). Randomizing who receives an intervention overcomes selection bias by trying to ensure that both the known and unknown characteristics of control and treatment groups are identical.[3]

Randomized experiments can be assessed according to the extent to which they adhere to internal and external validity criteria. Internal validity allows attribution of 'change' to the intervention in question, and is achieved through prior random assignment of the research sample to treatment and control groups. External validity allows findings to be extrapolated to a wider population, achieved through randomly selecting the research sample from a wider population.[4] Of the two criteria, randomized experiments comparative advantage is in internal validity. For example, both Deaton (2009) and Rodrik (2009) suggest that the (obsessive) control required for absolute internal validity compromises the ability of findings to be extrapolated to a wider population.

To give an overview of what randomized experiments consists of, we now describe five main stages in conducting a randomized trial (see Poverty Action Lab, 2008; Duflo and Kremer, 2005). This is a very crude summary, but gives a sense of what randomized experiments are about. A first stage is turning broad research questions into specific null hypotheses that the experiment is hoping to disprove (for example, that a given food supplementation program has no effect on recipients' growth or nutrition). This is followed by producing a theory-led assumed causal chain linking the intervention to the impact indicators in question (this doesn't have to be economic theory, although until now it appears to have been)

[3] Whilst this is clearly desirable, practitioners recognise that attaining this level of comparability across participant groups within a community or society is not straightforward (in contrast to experimental methods in the physical sciences) due to the existence of 'unobserved' or 'essential' heterogeneity (Heckman et al, 2006 in Ravallion, 2009a).

[4] Importantly, this may not be possible when the treatment group has specific characteristics, e.g. extreme poverty, evidence of child malnutrition.

and selecting key impact variables by which the null hypothesis will be judged (in this case, anthropometric measures, or clinical evidence of malnutrition observed through direct measurement). Importantly, the way in which these indicators are analyzed also needs to be set out to avoid post-hoc manipulation of data: typically, the mean effect on the group designated as the treatment group (known as the average effect on the treated).

Second, to select a sample size that is large enough to ensure that comparisons between treatment and control groups will be statistically significant and able to show a visible 'change', but is within the study's budget. Third, to randomly select treatment and control groups (which can be achieved through the use of public lotteries).[5] Randomizing can include the creation of multiple treatment groups to assess different components of interventions. Fourth, to collect data before and after the intervention, including piloting the research instrument, checks on data entry, cleaning data, etc. And fifth, data analysis where the mean figures of key impact variables for treatment and controls are compared. Confidence in the results depends on size of sample, the hypothesis, and the standard deviation of the outcome variables. It also depends on a range of checks having been conducted, drawn from existing good practice in medical randomized experiments. For example, checking that standard errors have been appropriately calculated and refraining from the use of covariates (Deaton, 2009, p36).

2.1 The strengths and shortcomings of randomized experiments

Randomized experiments are a powerful tool and have five key strengths (Banerjee and Duflo, 2008). First, a clear attempt to identify the effects of a specific (series of) intervention (s) (in other words, internal validity). Second, and as mentioned, an ability to offer answers to multiple hypotheses, through the creation of multiple treatment groups achieved through varying components of programs, or the sequence of interventions. Third, experiments can create a long-term relationship between evaluators (which until now have mainly been econometricians and their research students) and implementing agencies (such as donors or NGOs).[6] Fourth, results from randomized experiments are easy to convey, and often resonate well with policymakers and funding agencies. And fifth, randomized experiments can provide a basis for cost-benefit analysis.

But just as it is important to be open and realistic about the strengths of randomized experiments, we also have to be explicit and clear about their shortcomings (which until recently have not been discussed with enough candor – see Ravallion, 2009b). One shortcoming has been the selection of interventions evaluated through randomized designs. For example, Jones et al (2008) suggest there are significant gaps in the application of counterfactual impact evaluations (encompassing both randomized experiments and *ex post* quasi-experimental approaches). In particular, they highlight the lack of studies on environmental protection, agriculture and on gender issues. Woolcock (2009:6-7) relates the gaps in experimental evaluation to a similar bias that prevails in project funding:

[5] Alphabetisation may introduce bias if resources are allocated alphabetically because that is how many lists are presented.

[6] This could be less likely when projects are scaled up and implemented by national governments.

> Directors are going to have a much easier time being persuaded that funds given to build roads, enhance irrigation and immunize children will produce positive, measurable and immediate impacts, certainly when competitors for these same funds are proposing to address land reform, consolidate peace accords, or initiate efforts to improve the judiciary in 'failed' states [where] the metrics of success are inherently unclear.

This line of argument reflects the widespread belief that randomized experiments can only "can take only a very specialized type of evidence as input and special forms of conclusion as output" (Cartwright, 2007). There are also further reasons why randomized experiments are good for addressing certain research questions and not others. Experiments require time to ensure that interventions are embedded before the final research tool is conducted, and this may conflict with the short-term policy horizons of governments and donors (see Goldin et al, 2007). In addition, whilst randomized experiments are suited to small-scale development projects, they are not suitable for evaluating broad policy changes. For example, public sector reforms or changes to exchange rates or trade regimes are not appropriate due to the difficulty in establishing the counterfactual (see Goldin et al 2007; and Bhagwati, 2007). White (2007a:7) comments dryly that it is usually "not possible to randomly place large-scale infrastructure, such as a port or major bridge". Moreover, we should not forget political concerns: those with vested interests in a program (perhaps local political elites, or even donor or project staff) may have reasons to try and prevent a randomized evaluation (and prefer the *status quo* where procedures and impacts are opaque) (see Scott, 1998; Moore, 2007; and Bhagwati, 2007).

Putting aside broad questions about the applicability of randomized experiments to specific research areas to one side, there are also numerous limitations to randomized experiments within the research design itself, as acknowledged by practitioners (see Poverty Action Lab, 2008; Duflo and Kremer, 2005). We now outline six limitations related to internal validity, before turning to external validity.

Internal validity

First, attrition from samples, possibly as a result of the intervention or evaluation. This is shared by all types of longitudinal research, and can be partly overcome by tracking people if they move or if the household splits (although this is inevitably costly). Mortality cannot be overcome.

Second, the merging of treatment and control groups. In other words, when a control unit forces itself into the treatment group (for a variety of reasons, such as local or institutional politics). This poses considerable challenges for data analysis. Third, experimental designs can also suffer from spillover effects between treatment and control groups. For example, when the direct or indirect effects of the interventions leaks over from the treatment group into the control (such as when an agricultural intervention also increases labor demand in neighboring communities). Leakage can be mitigated through randomization procedures – for example, increasing the geographical distance over which control and treatment are selected (although a downside of this approach is that increasing the distance might reduce the geographical similarity of the two groups).

Fourth, a lack of compliance by an implementing agency. For example, the institution may not adhere to certain criteria, such as ensuring the separation of treatment and control groups, thus compromising the study. It is not unreasonable to expect some evaluators who are unused to experimental approaches to maintain the control required for randomized experiments.

Fifth, limited attention to sub groups. The conventional output from an experiment is the ATET (i.e. the average effect of the treatment on the 'treated'), rather than any one participant. Sub groups are often not reported. For this reason experimental findings tend to cloak the losses of those who might not have benefited from an intervention. For example, Deaton (2009, p29) states that "the trial might reveal an average positive effect although nearly all of the population is hurt with a few receiving very large benefits" and cautions that "much of the disagreement about development policy is driven by differences of this kind". In this respect, impact evaluation subscribes to a utilitarian notion of improving aggregate wellbeing (in other words, the greatest good for the greatest number). This clearly conflicts with more rights-based perspectives that are concerned with ensuring that no-one should fall below minimum thresholds (as illustrated above).

And sixth, there may be strong moral and ethical concerns against using portions of a population as a control group. For example, the provision of basic services in health and education is a human right, and withholding such services from a portion of a population as a control group may be ethically unacceptable and may cause avoidable harm. Proponents of randomized experiments suggest this shortcoming can often be avoided. As it is rarely possible to make an intervention available to everyone who needs it immediately, a common approach is to utilize pilot schemes. Researchers can thus employ a 'pipeline approach' using communities or households that have been selected for project but not yet treated as the comparison group (thus avoiding selection bias). This is a persuasive argument. Moreover, others argue that what is really unethical is to continue to spend billions of dollars on ineffective interventions, and that randomized experiments can provide a better evidence base for targeting resources efficiently as long as conclusions are founded on a clear understanding of how the intervention works (White, personal communication).[7] Again, a powerful argument. But not, in our view, entirely convincing. Consider the following hypothetical example.

Assume that a pilot program chooses randomization to assess the impact of two supplementary feeding flours (containing different proportions of maize, soya, groundnut flours enriched with vitamins) in a remote rural area with a sedentary population, and that the trial will take place over three years (with the resource envelope allowing the generation of a panel dataset of twelve hundred households in three waves). The impact indicators are rates of stunting (height-for-age) and wasting (weight-for-height) and the study uses census data to generate a random sample from households containing at least one infant (say, about fifty percent of households). Assume also that in the second year the region selected for the trial suffers from a rainfall shortage, staple grain prices spike by a factor of four, agricultural wage levels plummet (along with livestock and the prices of other assets). Broadly speaking, those children at greatest risk of permanent loss of stature, and cognitive ability, from the shock are

[7] C.f. the Proempleo scheme in Argentina (Galasso et al, 2004 in Ravallion, 2009) where the reason why the scheme was successful was simpler and cheaper than might have been supposed.

those in the lowest income quintiles (for example, see Hoddinott, 2006, on the long-term effects of the Zimbabwean drought of the early 1990s). What effect does the design of the evaluation have on the long-term prospects of infants in this area?

A non-experimental approach which compared the two products without a control would ensure that infants in all twelve hundred households had a reduced chance of suffering long-term harm. However, an experimental evaluation would include four hundred households with infants who could have received the supplementary feeding flour, but were denied it as they were part of the control. These four hundred households would contain many more than four hundred infants and children who, in the experiment without a control group are likely to have benefited from the supplementary feeding.

Although this is an extreme (and simplified) example, and best practice within experimental design would recognize the dangers illustrated, the broader point still stands: that withholding resources from poor people who live in risky environments so that they can constitute a 'control' group can create avoidable harm. Or in other words, experimental methods need to ensure that withholding treatment from a control will not, in any way, cause individuals to fall below a minimum threshold that might have a lasting effect on their wellbeing. Experimental approaches need to be acutely aware of the full range of risks faced by participants (as whilst shocks are unexpected, they are not unusual), and to face ethical issues with the seriousness and sincerity they deserve.

External validity

There are also four main limits to external validity (see Poverty Action Lab, 2008; Duflo and Kremer, 2005; Banerjee and Duflo, 2008). The first of these is the influence of context on the intervention. For example, Deaton (2009, p43) warns that "an educational protocol that was successful when randomized across villages in India holds many things constant that would not be constant if the program were transported to Guatemala or Vietnam" (c.f. Attanasio 2003 in Woolcock, 2009). This question has two parts. On one hand, the influence of the socio-cultural and physical environment on the intervention. In other words, would interventions judged to be highly successful in a randomized experiment in a particular setting have the same effect if implemented in a different region or country? Whilst economists tend to believe that individuals respond to the same set of incentives in a uniform manner (and are thus likely to think in terms of closed systems), many other social scientists often perceive reality as being much more complex, acknowledging that the social world is an open system). An example from the physical science – that of a falling leaf – helps to elucidate this point (Baert, 1998). If the physical world were a closed system then, according to the law of gravity, one might expect a leaf to fall from a tree in a straight line. Instead, falling leaves are subject to a wide variety of forces, and their trajectories are highly varied and difficult to predict. This is not to say that the law of gravity doesn't hold. Of course it does. But, even in the physical sciences, closed systems are unusual. The significance of this example is that in the social world, which is certainly an open rather than a closed system, the method of extrapolating from one context to another may not be as accurate as we would like. The second part of the context question is whether the implementing institution would be at all similar if it were scaled up. As Deaton (2009) notes, "small development projects that help a few villagers or a few villages may not attract the attention of corrupt public officials […] yet they would do so as soon as any attempt were made to scale up" (Deaton, 2009:44) And that "scientists who run the experiments are likely to do so

more carefully and conscientiously than would the bureaucrats in charge of a full-scale operation" (ibid). [8]

The second main limit to external validity is that interventions can cause changes in behavior that wouldn't occur if scaled up (for example, increased uptake at a pilot stage due to the novelty of the intervention). Third, that the evaluation itself causes the treatment and/or control groups to change behavior (in the literature these are known as the Hawthorne effect in the treatment group, or the John Henry effect in the control group – where people in the control group view themselves as being in competition with the treatment group and so change their behavior). This also reflects a concern among evaluators (e.g. Adato, 2007) that randomized experiments can increase social differentiation and even create conflict between beneficiaries and non-beneficiaries.

And fourth, equilibrium effects if scaled up (see Chen, Mu and Ravallion, 2006). A good example comes from Banerjee and Duflo (2008): an evaluation might find that extra-curricular tuition for lagging students improves employability post-education. However, if this was scaled up at a national level, the extra supply of school leavers who benefited from this tuition would limit each student's chances of getting a job.

Even when internal and external validity issues are mitigated to the greatest extent possible, some scholars are skeptical about the extent to which randomized experiments can generate 'gold standard' data. As suggested earlier, a "familiar trade-off between internal and external validity" as the formal methodology puts severe constraints on the assumptions a target population must meet to justify extrapolating a conclusion outwards from the treatment group (see Cartwright, 2007, p.11). And Deaton (2009, p6) concurs: "the price for this success [in internal validity] is a focus that is too narrow to tell us 'what works' in development, to design policy, or to advance scientific knowledge about development processes".

As we can see, randomized experiments suffer from a number of broad shortcomings and a range of more specific risks to internal and external validity, some of which can be overcome. A further shortcoming, until now, has been the limited use of qualitative methods.

3. MIXED METHODS WITHIN AN EXPERIMENTAL DESIGN

Randomized experiments have so far been dominated by quantitative methods, almost exclusively based on the survey instrument. For example, it is rare to see skilled and time-intensive methods such as ethnography used as part of a randomized experiment (although embedding anthropologists within institutions conducting randomized experiments would be highly beneficial). The dominance of quantitative methods is hardly surprising: the experimental methodology adheres to positivist principles and is very good at tackling *what* and *where* questions (which means it is good at capturing a state or condition). But, by relying only on quantitative methods, randomized experiments are often unable to tell us very much about

[8] See also Woolcock (2009:8) who highlights the example of the Kecamatan Development Project, Indonesia, which became more successful on scaling up as it learnt from its experiences and was able to attract better quality staff.

how or *why* societal change occurs – they often cannot inform us about key transmission mechanisms and therefore how interventions can or cannot be scaled up or transferred to other settings (see Thorbecke, 2007; Prowse, 2007). Adato (2007:9-10) notes for example that survey methods are at a disadvantage when it comes to unpacking the 'black box' of impact due to:

The necessary brevity of questions and the use of proxies that are often blunt measures; respondents' inability to sufficiently express what they mean in selecting among categorical or continuous variables; the limited ability of enumerators to follow up when more information or clarification is needed; and the difficulty of establishing the rapport and trust needed to maximize truthfulness in replies

Qualitative methods, on the other hand, are generally able to shed light on 'why' and 'how' questions, are good at capturing processes, and pay greater attention to who benefits from an intervention and who does not. Examples of where qualitative methods have been used in experimental designs in developing countries include Copestake et al, 2004 (microfinance); Rao and Ibanez, 2005 (social funds in Jamaica); Adato 2007, 2008 (CCTs in Nicaragua and Turkey); Alzua et al., 2007 (training program for disadvantaged youth in Latin America and the Caribbean); and Gibson and Woolcock, 2008 (increasing accessibility of legal mechanisms to poor women in Indonesia).

The following examples from White (2008) and Adato (2007) show how iterating qualitative and quantitative data within a counterfactual design (both randomized experiments and *ex post* experimental designs) highlights tensions and mismatches which, once overcome, vastly improves research findings. White's (2008) study of education reform in Ghana is a good example of the value of utilizing a mixed methods approach. At first, White's (2008) initial understanding of the subject was strongly influenced by early interviews with middle-class informants, and the insights of members of the research team (themselves part of the middle class). However, his initial conclusion that the US$300 million the World Bank had invested in basic education had produced little impact due to weak management structures was revised after reviewing survey data that showed large increases in enrolments, and a large decrease in illiteracy amongst primary school leavers. The survey data also showed the importance of Bank-funded classrooms and textbook to educational outcomes. A couple of days 'development tourism' provided White with further important insights: increasing differentiation amongst publicly-funded schools, exacerbated by a shift towards decentralization through community and district funding. Moreover, such fieldtrips showed how poorer communities were not able to supplement the basic infrastructure (a concrete floor, steel girder uprights, and a metal roof) provided by the Bank, even though this was a key project expectation. These insights were evident in the quantitative data: poor-quality classrooms led to poor learning outcomes, and children in the poorest regions were increasingly being left behind. The example illustrates how when rigorous qualitative research, including fieldwork visits, is sequenced appropriately with quantitative methods, it can provide insights that statistical analysis on its own might struggle to produce.

A further example of how qualitative methods can guide the interrogation of quantitative datasets comes from White and Masset's (2007) study of the Bangladesh Integrated Nutrition Project (BINP) in rural Bangladesh. BINP monitored the weight and height of children from birth to 24 months, and encouraged the mothers of malnourished or stunted children to attend nutritional counseling. Based on the anthropological literature regarding intra-

household decision making, White and Masset (2007) conducted participatory research and focus groups discussions to assess the extent to which knowledge about nutrition was put into practice by mothers. These qualitative methods confirmed findings in the anthropological literature: that a wife living with their mother-in-law had limited leverage over food issues. In this respect, by targeting a child's mother, and not their paternal grandmother, the BINP might be directing counseling at the wrong individual in the household. The qualitative findings were reflected in the quantitative data: married women with children who lived with their mothers-in-law had much less influence over food purchases and preparation. Moreover, the quantitative analysis found that in conservative rural areas the ability of such daughters-in-law to participate in nutrition counseling was severely circumscribed (a further reflection of their limited agency and power).

A final example comes from Adato (2007:17-20) who recounts how qualitative research in Nicaragua was used to identify two unexpected side-effects from the CCTs: firstly, that household targeting was creating new types of social differentiation among school children; and secondly, that children were being force-fed immediately prior to weighing to ensure they met the conditions for remaining in the program. The research also explained one paradox, namely why iron supplementation failed to reduce anemia in young children (because the iron supplements were diverted to older siblings).

These examples illustrate that restricting experimental designs to solely quantitative data may hide much more than it illuminates, and there is a very strong case for combining both qualitative and quantitative methods within studies. However, where qualitative methods are mentioned within an experimental design, they are often equated with offering context (e.g. through field visits) or participatory approaches (for example, see White 2008; Duflo and Chattopadhyay, 2004; Karlan, 2009). As yet there has been little explicit consideration of the extent to which qualitative or participatory methods might be able to be the *primary* measurement tool within a randomized design (despite the example of Duflo and Chattopadhyay, 2004, and that this is practiced within social policy).

A first step in such a process is to review the main categories of qualitative method. Here we outline five we have some familiarity with.

- Ethnography (or in other words, participant observation over a relatively long timescale)
- Semi-structured interviews (where the interview is guided by a checklist of pre-determined but open-ended questions)
- Life history interviews (there are numerous forms of biographical methods – here we refer to eliciting a respondent's life story and using this data to co-create a timeline for the respondent to discuss and interpret)
- Focus group discussions
- Task-based group methods, often used as part of 'Participatory Poverty Assessments', such as community mapping and ranking exercises

Whilst these approaches are generally seen as *qualitative* methods, this is not to say that they only generate qualitative data (Chambers, 2007). The qual/quant divide is really much more of a continuum. Instead of using the terms quantitative and qualitative, Kanbur (2001) highlights how social research methods sit at some point along five broad scales:

- Type of Information on Population: Non-Numerical to Numerical
- Type of Population Coverage: Specific to General
- Type of Population Involvement: Active to Passive
- Type of Inference Methodology: Inductive to Deductive
- Type of Value Framework: Multi-dimensional value vs. money-metric value

In the majority of cases, randomized experiments can be found at one end of the continuum, at least in the ideal case (numerical information, specific coverage, passive involvement, deductive methodology, and one-dimensional value framework). And frequently, qualitative studies are found towards the other extreme. But these methods can be combined to bring out the benefits of both traditions. For example, through 'converging' the methods so that qualitative methods take on the properties normally associated with surveys (see Booth, 2001, and Rao, 2001, who distinguish four different ways of integrating survey-based and qualitative research methods).[9] But to what extent will qualitative methods compromise the internal and external validity criteria of a randomized experimental design?

The following two figures show two matrices. The first compares the basic characteristics of randomized experiments with the five types of qualitative research. Each of the qualitative methods are assessed according to the likelihood that they could adhere to the basic characteristics of randomized experiments. Three simple categories are used: likely (light green), unclear (yellow) and unlikely (orange).

[9] Booth (2001) and Rao (2001) highlight four ways of combining qualitative and quantitative methods: (i) parallel - where the research methods are conducting separately and both inform the findings and outputs of the research; (ii) linkage - where contextual investigations, such as qualitative interviews, are a subcomponent of a sample survey, with the interviews fitted to survey sampling frames; (iii) convergence - where contextual methods take on properties normally associated with surveys (i.e. random sampling); (iv) triangulation - where different data sources, both between and within the two main methodological traditions, are sequenced and combined within the research design.

Table 1 – To what extent might qualitative methods adhere to the basic characteristics of randomized experiments?

	Ethnography (participant observation)	Semi-structured interviews	Life history interviews	Focus group discussions	Task-based group methods
Ex ante null hypothesis to be disproved					
Specified causal pathway					
Specified main variables					
Sufficient sample size for data saturation					
Randomly select treatment and control groups					
Research waves before and after intervention					
Data analysis					
Potential for use as primary research tool in experimental design					

Categories: likely (green - horizontal) / unlikely (orange - vertical).

The matrix suggests that one method appears unsuitable at this point – ethnography – mainly because of its attention to detail and deep immersion in circumscribed locations. It also reflects the inductive nature of ethnography, where research questions emerge from long-term participation and observation in a community and are usually not clearly defined prior to entering the field. This is not to say that ethnography couldn't run parallel to the main experimental design, or be used in a mixed methods design (Adato, 2007, 2008), but that the ethos of ethnography (not to mention the practicalities and cost) militate against using this methods as the *primary* measurement tool. The same argument applies to genuinely participatory research (e.g. participatory learning and action, PLA), which tend not to have an *ex ante* hypothesis, a predicted causal chain or *ex ante* selection of main variables due to an inductive and iterative approach to generating research questions (not included in Table 1). However, participatory methods, such as task-based group approaches, can be used within an experimental design (as illustrated by Duflo and Chattopadhyay, 2004), as such methods are increasing being used to generate statistics (see Barahona and Levy, 2003; Chambers, 2007) not least as part of participatory impact assessment approaches (e.g. Catley et al, 2008).

This leaves us with four possible methods: semi-structured interviews; life history interviews; focus group discussions; and task-based group methods. These four methods are

now compared in terms of the extent to which they compromise the internal and external validity of a randomized experimental design. This comparison is in terms of whether the qualitative methods might do better (green), the same (yellow), or worse (orange), than the conventional survey method.

Table 2 – To what extent do qualitative methods compromise the internal and external validity of randomized experiments?

		Semi-structured interviews	Life history interviews	Focus group discussions	Task-based group methods
Internal Validity	Attrition				
	Merging of treatment and control groups				
	Spillover effects				
	No institutional compliance				
	No sub groups				
	Moral or ethical concerns				
External Validity	Context - environmental				
	Context - institutional				
	Pilot creates effects				
	Evaluation changes behaviour				
	Equilibrium effects				
Cost					

Categories: Better (green - horizontal) / similar (yellow – diagonal) / worse (orange - vertical).

Table 2 suggests that focus group discussions and task-based group methods may do worse than the survey method in terms of spillover effects and the evaluation changing behavior, due to the open, public nature of these methods. For example, people may be reluctant to admit to receiving benefits from other sources or to not having changed their

behavior in the intended direction. That said, there is no reason to suppose that respondents will reveal this information to an official enumerator they have only just met, and it may be that free discussion within a focus group will give them more confidence to speak frankly. Overall, though, we feel that 'collective' methods will probably perform worse than individual methods. Group methods may also be more expensive, due to higher fixed costs per research encounter (although there may be a trade-off in terms of the numbers required for data saturation, especially within a clustered research design).

On the other hand, semi-structured and life history interviews do not appear to compromise the experimental design to any greater extent than the conventional survey methods. After all, a survey is typically based on a participant's responses in a one-on-one interview and the quality of the data depends on the quality of the rapport between the enumerator and the participant. In this respect, it can be argued that the dialogic nature of semi-structured and life history interviews will improve the quality of data generated, reduce the likelihood of attrition from samples, and can explore why an individual's actions might be altered due to the evaluation and assuage fears and rumors. This brings us to two broader points. First, these qualitative methods, by their nature, are also likely to perform better than the survey tool in understanding contextual threats to the experimental design. This is in terms of both the influence of the socio-cultural and physical environment on the intervention, and whether institutions will act differently if the intervention is scaled up. Whilst this clearly has implications for the piloting of measurement tools (in other words, that using a qualitative method within the piloting phase could highlight potential threats), it also has implications for using qualitative methods as the *primary* measurement tool. For example, qualitative methods can help to explicate how aspects of a local environment (whether political, social or physical) might be idiosyncratic, and can capture institutional peculiarities and possible dysfunctionality to a much greater extent than the survey method (e.g. Gibson and Woolcock, 2008).

And second, qualitative methods are also much more likely to tell us *why* an intervention succeeds or fails compared to the survey method. For example, Ahmed et al.'s study of a conditional cash transfer in Turkey (2006, in Adato, 2007, p22) demonstrated that the reluctance to send daughters to secondary schools went beyond schooling costs as "secondary schools are often far from home, and transportation options are not trustworthy with respect to [girls'] honor". So, even though the CCT alleviated the burden of school expenses and prevailing poverty "where the other factors were strong, the cash could not compensate" (ibid). Qualitative methods can tell us about the importance of such key transmission mechanisms and societal norms. In sum, using qualitative methods as the primary measurement tool not only adds contextual explanation to the average treatment effect on the treated (ATET), but can offer a much richer and more accurate approximation of causal processes than solely using a survey measurement tool.

This penultimate section now discusses which of these two methods might be best suited for experimental designs? In other words, if a funding agency wanted to allocate scarce resources to conduct randomized evaluations using a qualitative method, which method might be first in line? In our opinion, it could well be life history interviews. Why? There are three reasons.

First, the longitudinal focus of a life history interview resonates with the 'before and after' criteria of 'double difference' experimental designs. Second, a life history interview

highlights the importance of social relations and institutions for assessing the intervention in question (birth, childhood, school, marriage, children, employment perhaps). And third, life history interviews allow the generation of quantitative, qualitative and visual data.

But that is not to say using life history interviews within an experimental design doesn't have a number of shortcomings. For example, the cost per interview will be higher (due to the greater duration per research encounter, and the fewer numbers of interview per day), expanding the resources required for the study, or reducing the power of the findings. The training of researchers will also be more expensive, as few have experience of conducting this form of research method. Using this retrospective dialogic method also raises ethical concerns: asking individuals to recount the trajectory of their life often brings painful memories to the surface (particularly in developing countries where citizens endure much greater levels of risk). Will researchers be able to disengage from respondents in an ethically acceptable manner? Moreover, generating large amounts of qualitative and visual data presents an interesting challenge in terms of analysis and interpretation.[10]

Whilst these shortcomings are important, they could be overcome. A good template of how life histories could be the primary tool within a randomized evaluation is provided by an on-going poverty dynamics study by Davis and Baulch in Bangladesh (see Davis and Baulch, 2009). This study combines a quantitative panel survey of 1787 households with a sub-sample of around 300 qualitative life history interviews, all of which generated visual trajectories.

4. CONCLUSION

All methodologies have limitations. Experimental design is a valuable option (with due consideration of applicability, threats and ethics) within the spectrum available to researchers and evaluators, particularly when qualitative methods are included within the methodology. For example, Woolcock (2009:13) views this as the factor that moves a methodology from 'gold' to 'diamond' standard. Mixing methods within an experimental design may reduce the need for speculative interpretation of quantitative results, avoid fundamental misunderstandings due to neglect of the context in which the intervention is taking place, foster greater engagement with evaluation communities and, more importantly, with the beneficiaries of interventions. But, as yet, there appears to be little appreciation that just because randomized experiments utilize a relatively strict positivist methodology, this doesn't preclude qualitative methods from taking an equal or primary role as the data measurement tool. The next steps in advocating for a greater number of experimental studies that utilize a qualitative method as the primary measurement tool are to: (i) assess the implications of using qualitative methods in terms of the skills of research personnel, and institutional acceptance; (ii) conduct a detailed comparison of the interview-level strengths and shortcomings of different measurement tools within the rubric of a randomized design. From our perspective, moving this research agenda forward chimes with Banerjee and Duflo's (2008) call for 'creative experimentalism', and may

[10] The authors would like to thank Peter Davis for raising a number of these issues.

help to bridge the gap between help advocates of randomized control trials and mainstream evaluation communities (see Leeuw and Vaessen, 2009).

REFERENCES

Adato, M. (2007) Combining Survey and Ethnographic Methods to Evaluate Conditional Cash Transfer Programs. Q-Squared Working Paper No. 40.

Adato, M. (2008). “Combining survey and ethnographic methods to improve evaluation of conditional cash transfers” in In P. Shaffer, R. Kanbur, N. Thang and E. Bortei-Doku (eds), Special Issue Journal of Multiple Research Approaches (Vol. 2, No. 2).

Asian Development Bank (ADB) (2006) Impact Evaluation: Methodological and Operational Issues. Manila, Philippines: ADB.

Baert, P. (1998) Social Theory in the Twentieth Century. Cambridge: Polity Press.

Banerjee, A.V. et al (2007). Making Aid Work. Cambridge, MA: MIT Press.

Banerjee, A., S. Cole, E. Duflo, and L. Linden (2006). ‘Remedying Education: Evidence from Two Randomized Experiments in India’. CEPR Discussion Paper 5446. London: Centre for Economic Policy and Research.

Barahona, C., Levy, S. (2003). How to generate statistics and influence policy using participatory methods in research: reflections on work in Malawi 1999–2002. IDS Working Paper 212.

Bhagwati, J. (2007) 'If it is Hard to Think of Aid Being Spent Productively in Africa, Why Not Spend Elsewhere for Africa?', in Banerjee et al.

Booth, D. (2001). Towards A Better Combination Of The Quantitative And The Qualitative. Q-Squared: A Commentary On Qualitative And Quantitative Poverty Appraisal. R. Kanbur (Ed), Cornell University Department of Applied Economics and Management Working Paper No. 105.

Cartwright, N. (2007). Are RCTs the gold standard? BioSocieties, 2, 11–20.

Catley, A., Burns, J., Abebe, D., Suji, O. (2008). Participatory Impact Assessment A Guide for Practitioners. Feinstein International Centre, Tufts University.

Chambers, R. (2007) Who Counts? The Quiet Revolution of Participation and Numbers. IDS Working Paper 296, IDS, University of Sussex, Brighton.

Chen, Shaohua, Ren Mu and Martin Ravallion (2008) “Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China,” World Bank Policy Research Working Paper No. 4084. March.

Copestake, J., Johnson, S., Wright, K. (2004). ‘Impact assessment of microfinance: towards a new protocol for collection and analysis of qualitative data’, in J Holland and J Campbell (eds) Methods, knowledge and power: combining quantitative and qualitative development research, Swansea: IT Publications, with Centre for Development Studies, University of Swansea.

Davis, P., Baulch, B. (2009). Parallel Realities: Exploring Poverty Dynamics using Mixed Methods in Rural Bangladesh. Paper presented at 'Escaping Poverty Traps: Connecting the Chronically Poor to Economic Growth' conference, Washington DC, February 26-27, 2009.

Deaton, A. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. The Keynes Lecture, British Academy, October 9th, 2008.

Duflo, E and Chattopadhyay,R. (2004) "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India,", *Econometrica* 72(5): 1409-1443.

Duflo,E., Banerjee,A. (2008) 'The Experimental Approach to Development Economics' CEPR working paper No. DP7037, London: Centre for Economic Policy and Research

Duflo, Esther and Michael Kremer (2005) "Use of Randomization in the Evaluation of Development Effectiveness," from George Pitman, Osvaldo Feinstein and Gregory Ingram, ed., Evaluating Development Effectiveness, New Brunwick: Transaction Publishers.

Easterly, W. (2006). The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good. New York: Penguin.

European Evaluation Society (EES) (2007) The Importance of a Methodologically Diverse Approach to Impact Evaluation – Specifically with Respect to Development Aid and Development Interventions. EES Statement. Nijkerk, Netherlands: EES secretariat.

Foresti, M. (2007) 'A Comparative Study of Evaluation Policies and Practices in Development Agencies'. Report for AFD Evaluation Department.

Gibson, C.M., Duncan, G.J. (2000). Qualitative/Quantitative Synergies in a Random-Assignment Program Evaluation. Paper presented at Discovering successful pathways in children's development: Mixed Methods in the Study of Childhood and Family Life Conference.

Goldin, I., Rogers, F. I. and Stern, N. (2007) 'We Must Tackle Development Problems at the Level of the Economy as a Whole', in Banerjee et al.

Hoddinott, J. (2006). 'Shocks and their Consequences Across and Within Households in Rural Zimbabwe'. *Journal of Development Studies*, 42 (2): 301-21.

Jones, N., Jones, H. Steer, L., Datta, A. (2009). Improving impact evaluation production and use. ODI Working Paper 300. Overseas Development Institute, London.

Kanbur, R. (ed.) (2001) Q-Squared: A Commentary On Qualitative And Quantitative Poverty Appraisal. Working Paper No. 105. Ithaca, NY: Cornell University Department of Applied Economics and Management.

Karlan,D. (2009) Cairo Evaluation Clinic: Thoughts on Randomized Trials For Evaluation of Development Economics Department Working Paper No. 65 / Economic Growth Center Discussion Paper No. 973, Department of Economics, Yale University.

Karlan, D. and Zinman,R. (2009) Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. *Review of Financial Studies*.

Leeuw,F. and J.Vaessen (2009) 'Impact Evaluation and Development: NONIE Guidance on Impact Evaluation' Network of Networks for Impact Evaluation (NONIE), World Bank, September 2009.

Lewis, J. (2007). Analysing Qualitative Longitudinal Research in Evaluations. Social Policy & Society 6:4, 545–556.

Lloyd-Sherlock, P., Locke, C. (2008). Vulnerable Relations: Life Course, Wellbeing and Social Exclusion in Buenos Aires, Argentina. Ageing & Society, 2008, 28:779–803.

London, A.S., Schwartz, S., Scott, E.K. (2005). Combining Quantitative and Qualitative Data in Welfare Policy Evaluations in the United States. Q-Squared Working Paper No. 12.

Molloy, D., Woodfield, K., Bacon, J. (2002). Longitudinal qualitative research approaches in evaluation studies. A study carried out on behalf of the Department for Work and Pensions. Working Paper 7, Social Research Unit, London.

Moore, M. (2007) 'The New Private Philanthropies Could Challenge the Existing Aid Business', in Banerjee et al.

Olken, Benjamin. 2007. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. Journal of Political Economy 115: 200-249.

Poverty Action Lab (2008) 'Evaluating Social Programs' Notes from short course conducted by J-PAL South Asia at the Institute for Financial Management and Research, Chennai, India, July 28th – August 1st, 2008.

Prowse, M. (2007) 'Aid Effectiveness: The Role of Qualitative Research in Impact Evaluation'. ODI Background Note, December.

Rao, V. (2001). Potters And Slums: Two Qualitative/Quantitative Projects In India. In Q-Squared: A Commentary On Qualitative And Quantitative Poverty Appraisal. R. Kanbur (Ed), Cornell University Department of Applied Economics and Management Working Paper No. 105.

Rao, V. and Ibáñez, A. M. (2005). 'The social impact of social funds in Jamaica: A mixed-methods analysis of participation, targeting and collective action in community driven development'. Journal of Development Studies 41(5), 788-838.

Ravallion, M. (2009a). Evaluation in the Practice of Development. The World Bank Research Observer.

Ravallion, M. (2009b). Should the Randomistas rule? Economists' Voice.

Rodrik, Dani. (2009). “The New Development Economics: We Shall Experiment, but How Shall We Learn?” In Jessica Cohen, and William Easterly, eds., What Works in Development? Thinking Big and Thinking Small, Washington: Brookings Institution Press.

Sachs, J. (2005). The End of Poverty. New York: Penguin.

Savedoff, W. D., Levine, R. and Birdsall, N. (2006) When Will We Ever Learn? Improving Lives Through Impact Evaluation. Report of the Evaluation Gap Working Group. Washington, DC: Center for Global Development.

Scott, J. (1998). Seeing Like a State: How Well-Intentioned Efforts to Improve the Human Condition Have Failed. New Haven, CT: Yale University Press.

Thorbecke, E. (2007). The Evolution of the Development Doctrine, 1950-2005. In G. Matrovas and T. Shorrocks (Eds.), Advancing Development: Core Themes in Global *Economics.* Basingstoke: Palgrave Macmillan.

White, H. (2008) ‘Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation’ NONIE Working Paper No. 6, World Bank, January 2008.

White, H. (2007a) ‘Technical Rigor Must Not Take Precedence Over Other Kinds of Valuable Lessons’, in Banerjee et al.

White, H. (2007b) ‘Evaluating Aid Impact: Approaches and Findings’. In S. Lahiri (ed.), The Theory and Practice of Foreign Aid, vol. 1. New York: Elsevier.

White, Howard and Masset, Edoardo (2007) ‘The Bangladesh Integrated Nutrition Program: Findings from an Impact Evaluation’, Journal of International Development 19: 627–52

Woolcock, M. (2009). Towards a Plurality of Methods in Project Evaluation: A Contextualised Approach to Understanding Impact Trajectories and Efficacy. BWPI Working Paper 73.

World Bank (2007) Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank. Washington, DC: Independent Evaluation Group, World Bank.

Worrall, J. (2007). Evidence in Medicine and Evidence-Based Medicine. Philosophy Compass 2/6: 981–1022

University
of Antwerp

IOB
INSTITUTE OF DEVELOPMENT
POLICY AND MANAGEMENT