Some thoughts on how to do research in development economics:

Applied (micro)theory, empirics and survey design

JM Baland, January 2025

Structure of the presentation

- These notes are for students at the MA/early Phd level, who are interested in developing a research project.

- We will discuss three topics:
  1. How to do an applied microeconomic model?
  2. How to do an empirical research?
  3. How to collect my own data?

Why do we read too much or not enough?

- In general, define first your **area of interest** and possible ideas you have on this.

- Read some but NOT all papers on the topic. Be aware of the last papers and more or less how they approach an issue. **Reading too much kills your creativity**: you have to design your research approach based on your own thinking. You have to think by yourself to add something that remains little explored.

- By contrast, read a lot of unrelated papers and construct your own general **economic 'culture'.** (This includes sociology, pol science, history, economic theory,…) This is critical for your creativity (eg Duflo on Indonesia). Read papers vaguely related to your question, both in theory and empirics, to enlarge the scope of your initial question. Do not hesitate to check on IA or in surveys for a summary of the possible arguments that are there.

- In your paper, you will refer to the literature, but instead of a list of all authors who wrote on the topic, you have to describe carefully what you add. You therefore have to select the papers to which you are most closely related, and discuss what this ‘relation’ is: why do you differ? What do you add? …

- You may also wish to attend seminars in development, in particular those proposed on line by the CEPR, the World Bank or BREAD.

- You need not do this alone. Co-authorship is a good way to confront your ideas, complement your skills and learn a lot (we all have a different way to think about the world). As a junior for a first research, seeking the (real) support of a more experienced researcher is a good way to start!! And co-authorship is one of the most interesting and rewarding human experience while doing research!

Part 1: How to do applied (micro) theory?

One paper, one idea…

- One paper, one idea… Theory often starts with common sense. A model is an abstract logical reasoning, aiming at describing a ‘mechanism’. We focus mostly on the causes and consequences of individual rational behaviour.

- You do development or environment, you must absolutely go beyond your field and listen to the other disciplines. In the interdisciplinary dialogue, we as economists are mostly interested for what matters on average, ‘regularities’, which in my view is our main difference with classic anthropology. We dont have much interest for ‘particularity’, unless they say something more general on human behaviour (tip of the iceberg).

- We tend to have also a bias towards what we can measure, even in theory.

How to do applied (micro-) theory?
Start with a question!

- Determine the main question or phenomenon you want to examine. The idea may come from your personal experience, a discussion with friends, readings of some theory or field visits. For example, you might want to understand "How do consumers respond to price changes in the smartphone market?" or "Why are they so few cooperatives managed by the workers?“
  - For instance, Gary Becker thought of a theory of rational crime while trying to park his car in new York just before a class, and deciding to ‘take the risk’ of a ticket.

- In this opening stage, you must **remain open**, let yourself be questioned: ‘this does not sound right’; ‘why do they do this? ‘; ‘it is not how I first thought about it’;….
  - For instance, I was struck by the insistence in economics on income inequality, while longevity constitutes another important metric of welfare, with very little work in economics. This led to a project with two colleagues on poverty measurement which explicitly account for life expectancy.

How to do applied (micro-) theory?
Start with a question!

- In development economics, these questions may come from **field visits** and participatory observations. These are very useful to evaluate the relevance of your questions, call into question your own pre-conceptions,…
  - For instance, some scholars started to interview polygamous wives, and from these interviews, came the idea that, in this context, the wives were in fact collaborating, and happy to do so, in colluding against their husband.

- Alternatively, you may find that a well-known model rests on a particular assumption that is very ‘restrictive’, in the sense that relaxing it offers a new perspective on the question (eg, moving from partial to general equilibrium, from static to dynamic, allowing some market imperfections or not, enlarging the set of feasible contracts, …). In other words, you have a counter-argument to some well-known result.
  - For instance, introducing dynamics in the static argument made by Martin Weitzman on the privatization of the commons.

How to do applied (micro-) theory?
Start with a question!

- Once you have the **intuition** for an argument or a counter-argument, start thinking about it using **basic logic**. This is the hard part, you have to be creative!
  - For instance, in the 'privatization of the commons' example, if private owners reduce catches today, it is to better (profitably) preserve the resource and have more catches tomorrow. So, total use of the commons must increase.

- Discuss your intuition and your arguments with friends or co-authors so as to feel the strengths and the weaknesses of your argument.

Why a model?

- The aim of the model is to test the **logical consistency** of this argument, but also **its ‘generality’,** in the sense of defining the largest ‘environment’ under which it holds, while keeping it as simple and convincing as possible. Unnecessary details are undesirable.

  - Your argument is not necessarily the ‘description’ of a mechanism. Sometimes, your argument aims to show the impossibility of something or the contradiction between two mechanisms in the same environment. The term ‘argument’ used here should be taken in a broad sense.

  - A model may also aim at bringing clarity in complex situations. You then want to disentangle different mechanisms, and their relative importance in different environments. Think of the Slutsky equation, for instance.

Start simple

Many important models are **‘not realistic’** and do not lend themselves to simple ‘falsification’ tests or direct empirical strategies, but help in our understanding of the world.

Take the Ricardian equivalence, the ‘law’ of comparative advantages or the Hotelling resource pricing rule:

$$P_{t+1} = P_t\,(1+r)$$

which is a fabulous intuition (exhaustible resources are to be considered as an asset, subject to returns) but is not directly testable. However, it remains a beautiful guide to our understanding of the economics of exhaustible resources.

Start simple

Simplify Reality : You have to set limits to keep the model manageable. Economic models are abstractions, so they require assumptions to reduce complexity.

Avoid 'realism' and irrelevant details and take the simplest set of assumptions needed for your argument. Typically, you start with a very simple 'example' that expresses the mechanism you had in mind.

- Focus on the **key variables and trade-offs or choices** that are relevant,
- For all aspects of the environment that do not matter (are 'orthogonal') for your argument (you may question this later), take the simplest assumption you can think of!

    For instance, we know that people are different, so that you may want to bring in heterogeneity in skills, in preferences,… But if this is irrelevant for your argument, just assume that people have identical preferences, skills,…

Example: a model of child labor

- Why would loving parents send their child to work, at the cost of their well being and future human capital?

- How do we think about this? Which assumptions do we make?

A model of child labor: what assumptions do we make?

- One parent, one child.
- Since we think of the child future (human capital), we need at least two periods: one where the child is a child, and one where he is now an adult.
- Each child has 1 unit of time, which he allocates to child labor or ‘education’: **1 = e+l**
- But parents love their children (this is an important assumption, of course), so that they care about them and their future: $\mathbf{U_P = U\,(c_p) + \alpha\, U_{child}}$
- Why would they make wrong decisions? Mainly because they are too poor now, and are unable to take money from their children future (as an adult) income. So we need capital market imperfections: we assume that parents cannot borrow!
- As you can see, these assumptions are not realistic (!) but we have to focus on the most important one, which is the capital market imperfection, and the substitution between e and l in the ‘time’ constraint.

And the discussion of one assumption by Dubois and Bommier 2004

“Baland and Robinson (2000) investigate the conditions under which decisions by parents about their own children’s work are inefficient. Using a simple two-period model with altruistically linked family members, they show that child labor decisions are efficient when credit markets are perfect and intergenerational altruistic transfers are nonzero. Moreover, they show that when the level of child labor is inefficient, because of liquidity constraints or because altruistic transfers are at a corner, a ban on child labor can be Pareto-improving.”

We argue here that the results of Baland and Robinson **are significantly altered when preferences account for the fact that children have a disutility for labor**. We find that child labor may be Pareto inefficiently high even if markets are perfect and there are altruistic transfers. This economic inefficiency is not related to market imperfections, but is a consequence of the noncooperative game in which altruistic parents are involved.”

Another simple model of child labor (Basu and Van)

- Intuition: child labor comes as a competitor to parental labor on the labor market. This must depress their wages in equilibrium, giving rise to multiple equilibria. It must be possible that, if all parents take their children out of the labor market, adult wages will increase enough for parents not to find it anymore worthwhile to send their child to work.

- Assume one parent, one child in a static framework
- Individual labour supply is either 0 or 1.
- Assume a competitive economy

- Main assumption: when parents are poor enough (below W), they send their children to work. If rich enough (above W), they do not. This generates multiple equilibria, and a ban on child labor may increase household welfare. (Basu and Van)

Make assumptions

1. Your assumptions should have strong **micro-foundations**.
2. Start with the fundamentals: utility and production functions, (or in rare cases, demand and supply). Assume that "consumers are rational" or "firms maximize profit“.

   (If they dont, you can show almost anything and run the risk of ad hoc theorization. The idea or rationality is a good way to start. If you have to work on non rational behavior, look at the guides in the literature that will help you to do that in a consistent manner)

3. How many agents do you need to make the point? How many goods and markets? How many decisions? How many states? Are you in a static of dynamic framework? If dynamic, do you work with a 2 period or with an infinite horizon model?
4. Do agents play strategically or in a competitive way (eg prices are exogenous)? What is the role of imperfect information (limited liability, moral hazard, adverse selection,…) and strategic behaviour?
5. Are you in partial or general equilibrium?

Make assumptions

- Be careful with **self-confirming assumptions.** This is one of the most difficult aspect of modelling: what you assume may in fact be the result of the model, not an assumption per se.
  - For instance, in ‘why are workers cooperatives so rare?’, we know that they exist among lawyers, small scale medical clinics, some high tech ventures,… You could be tempted to exclude by assumption these high value added service sector. This is precisely what you should not do: the fact that you have cooperatives in those sectors should be a result of your model, not an assumption.

It is not because we ‘observe’ something that this is a ‘good’ assumption to make. This observation may itself be the equilibrium result of a mechanism.

Start simple

- Choose **simple functional forms** that make sense in theory (i.e. are micro-founded).
    - utility functions: Cobb-Douglas (no income effects); quasi-linear (for a luxury good),…
    - Production functions: linear (one good), Cobb Douglas,… Assume constant returns to scale
    - Linear supply and demand functions.
    - Be aware of assumptions that guarantee unicity (if needed) or existence (eg Inada end point conditions).

- Be aware of the **simple tools** that help define a model:
    - In a contest problem, the contest function
    - With uncertainty, assume expected utility and Bayesian updating
    - For household decisions, start with the collective model
    - With bargaining, use the Nash solution
    - In a voting problem, start with the median voter or the citizen candidate,
    - In the production sector, start with price taking or monopoly pricing
    - …

Solve the simple Model

- Make a clear distinction between the **technical assumptions** that allow you, for instance, to present a closed form solution (eg Cobb Douglas utility, or a representative agent), and the **fundamental assumptions**, which determine the ‘world’ you are in.

- **Optimization**: find the optimal solution to the agents’ problem. These solutions need not be explicit, even though, to start with, an explicit solution may be easier to deal with.
- **Equilibrium**: Determine the model's equilibrium, where supply equals demand or marginal cost equals marginal revenue,…
- **Comparative Statics**
- **General equilibrium:** your analysis can be partial, but may also need to hold in a gen eq model. You then have to wonder or test whether what holds in partial also holds in general, allowing for feedback effects of the other markets.

Analyze and Interpret

- **Interpretation**: Translate mathematical results into **economic insights.** If relevant, explain why your result aligns, if relevant, with real-world observations.

- **Microfoundations** again: Once you converged on a good example that works, look at the literature on the assumptions underlying those ‘simple tools’ and perhaps choose a more appropriate one (consumption analysis (Deaton-Muellbauer); contest functions (Skaperdas); polarization and inequality (Esteban-Ray); expected utility or discounted utility flows (Barbera-Hammond) versus loss aversion or hyperbolic discounting (Loewenstein, Rabin, O’Donoghue,…),…

Think and work again

1. It is often the case that your initial question, once you start working seriously on it, changes, gets larger, has more implications than you thought, or does not hold in the environment you thought, but holds under another set of conditions. **Be open and prepared for surprises** and try hard to understand them. Surprises force you to think deeper, and the model is a help in that process. You have to understand the mechanisms behind the ‘surprise’: are they an artefact, or are they a fundamental consequence of the environment you describe?

Think and work again

- Having a **negative result** (‘it doesn’t work’) can also be taken positively, showing that a simple intuition does not work in a very simple environment designed to describe it. This may lead to ‘contradictory’ or ‘impossibility’ results, in the sense of ‘why this cannot be true’, which are also of great value.

- You may also develop a model to answer one question, and realize that it does not really work the way you thought, but that in the meantime your model is perfectly suited to answer another, related, question. This is one of the good ‘surprise’ of theory: it may lead you to issues that you did not think about beforehand (eg: my paper on child labor started as a paper on optimal fertility in a non–cooperative framework).

Converging on a ‘good’ model

Adjust Assumptions to

1. **test the validity** of your results in larger contexts. You must identify the assumptions that are the most critical for your result, and test the robustness of your argument. These assumptions are the logical conditions under which your ‘argument’ holds; in the sense of **‘even if…, this holds true’.** This is the key test for an ‘interesting’ model. You then have to ‘generalize’ those assumptions:
   - either by making them more abstract, more **general** (any increasing concave utility function, for instance, instead of the Cobb Douglas)
   - or by taking assumptions that a priori go **against your result** (eg, why do people make gifts if they are completely egocentric?)

Converging on a ‘good’ model

Adjust Assumptions to

2. **improve the relevance** to the question raised, provided this does not make your model too complicated. (For instance, if assuming "no transaction costs” is too restrictive in the context you are interested in, you could incorporate small transaction costs.) You have to try several times, with different sets of assumptions (that make sense) till you are more or less satisfied and understand your example better. This is an iterative process, with a lot of ‘versions’…

3. If relevant, develop **empirical implications.** Your results should provide ways to ‘cut the data’ under the form of possible correlations, heterogenous effects or causal implications that you can possibly illustrate with data. For an ‘applied theory’ paper, these illustrations really matter…

Converging on a ‘good’ model

- **Concision, simplicity and elegance** are required. The simpler and elegant it is, the easier it is to communicate and the more convincing you will be. Use graphs or simple examples.

- **Seek feedback**, share your model and your ideas, to reveal potential improvements or alternative approaches. Model-building in microeconomics is an iterative process that takes several new versions to be satisfactory.

A ‘good’ model…

The value of your contribution is all the better:

1. The more ‘general’ are your assumptions

2. The more ‘relevant’ is your question: ‘Why do we care?’

3. The more ‘elegant’ and simple is your model

4. The more ‘non-trivial’ or unexpected are your results.

A ‘good’ model…

The more ‘non-trivial’ or unexpected are your results. Examples are:

- Adam Smith-Arrow-Debreu: Can selfish behaviour in a competitive society lead to an optimal outcome?
- Akerlof: Can markets disappear if the seller is better informed than the buyer?
- Solow: Is infinite growth possible by only accumulating capital?
- Besley et al: Why does affirmative action may be self-defeating if it re-inforces negative stereotypes?
- Ricardo: can mutual trade be beneficial, even if one party is systematically (‘absolutely’) less productive in all activities or sectors?
- …

Train yourself

- Do not hesitate in taking well known models in the literature, typically topfield journals, and wonder what happens if you change some of the assumptions there. This is a good way to train with different types of proofs, gain confidence and better understand the role of assumptions in a model.

Part 2: How to do empirical research?

Five Key Characteristics of a Good Empirical Model and ten commandments

1. **Strong Theoretical Foundation**: The model should be rooted in economic theory, guiding the choice of variables, functional forms, and relationships.
2. **Data Quality**: A good model requires reliable and sufficient data.
3. **Identification Strategy**: The model should carefully distinguish causation from correlation using methods such as fixed effects, instrumental variables, or randomized controlled trials.

4. **Simplicity and Parsimony**: It should be as simple as possible but complex enough to capture key relationships. Parsimony minimizes unnecessary parameters, avoiding overfitting and allows communication.
5. **Accuracy and Robustness**: It should predict or explain the data well and should not rely on one particular estimation on a particular sample.

1. Develop a Clear, Focused Research Question based in Theory

- **Theoretical Motivation**: Ensure your question is grounded in economic theory or has a well-defined hypothesis, which will guide your empirical design and provide a basis for interpreting results. Why?

- Best case: field observations → theory → empirical validation!

Why theory (and common sense!)?

1. **To guide Hypotheses and Research Questions**

- Economic theory helps define what we expect to observe, i.e. **testable hypotheses**.
    - For instance, if we use labor market theory, we might hypothesize that minimum wage increases could impact employment levels.

- **Avoiding Data Mining**: Without theory, researchers might run analyses without clear purpose, leading to "data mining" or spurious results. Theory prevents the risk of generating patterns that appear significant but lack real-world meaning.

Why theory?

2. Causal Relationships can only be based on some ‘theory’

Direction of Causality: Theory suggests the causal direction and helps interpret relationships between variables.

- For example, economic theory on consumption suggests income changes impact spending, rather than vice versa, giving a logical foundation.
- Economic theory on utility maximization, profit maximization, present bias or risk aversion informs empirical models and help us to specify relationships and causality correctly.

Why theory?

3. To indicate the relevant Variables and Controls and how to measure them

Selecting Variables: Theory helps identify which variables can influence the outcome, either as the ‘main’ variables of interest, or as a control.

- For example, demand theory suggests controlling for income when examining how price affects quantity demanded, helping to isolate the true price effect.

Why theory?

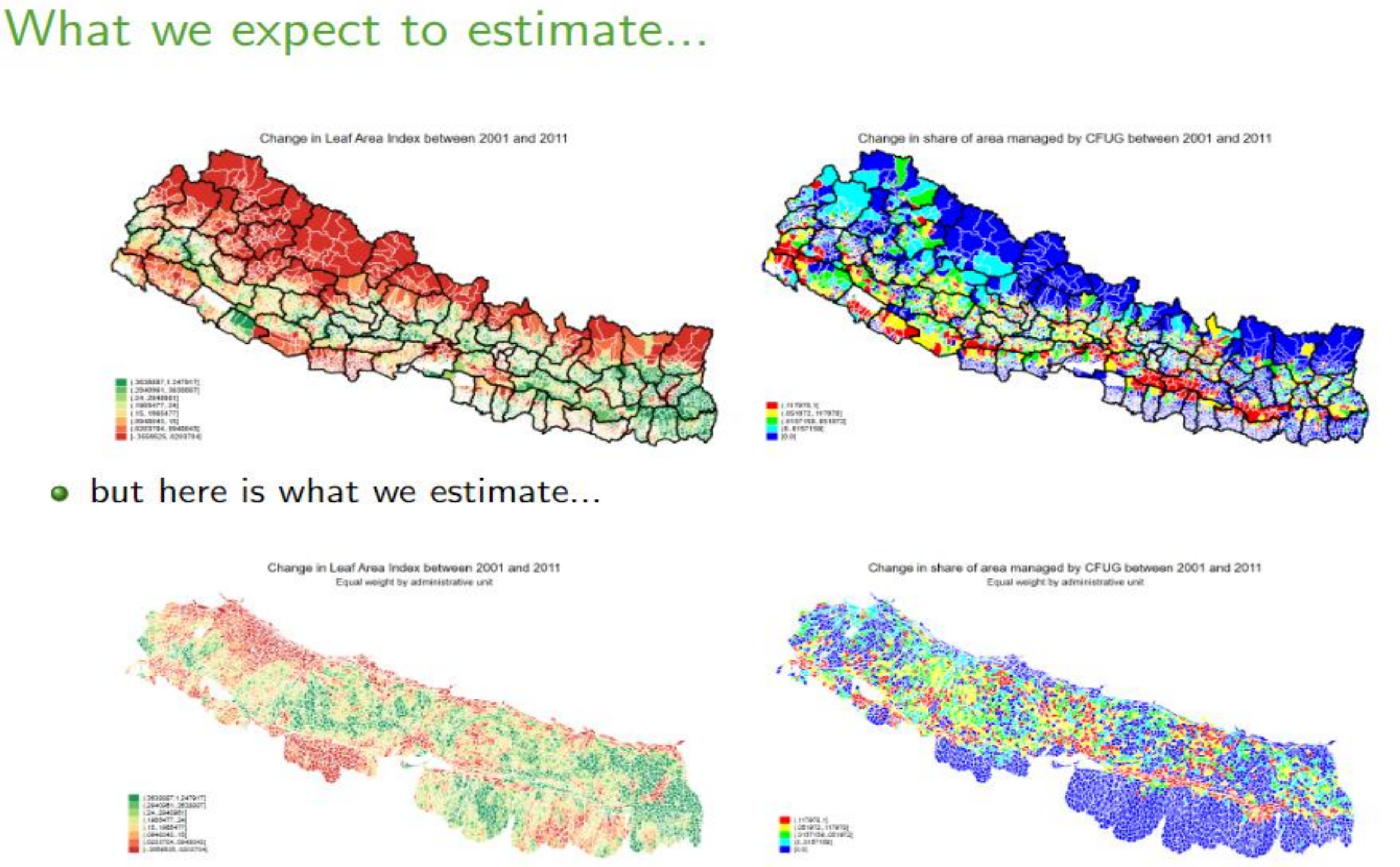
3.1 Theory indicates the unit of observation, i.e. the level at which to collect information

Some thinking in general is usually enough. Making sense is ‘theory’.

> For instance, suppose that your interested by the impact of community management on forest in Nepal. Given that your outcome is based on a surface (forest area), the level of information to be used is the unit area (ie the pixel in satellite imagery). Suppose that you are interested in how government directives translate to community management: the level of information to be used is the Village level (VDC). Suppose finally that you are interested in how community management of forests affect households: the level of information to be used is the household.

Fundamentally, this is a question of weighting and clustering, but this is deeply grounded in (simple) theory.

Measuring biomass in Nepal: an example (F. Libois 2024)



What we expect to estimate...

Change in Leaf Area Index between 2001 and 2011

Change in share of area managed by CFUG between 2001 and 2011

- but here is what we estimate...

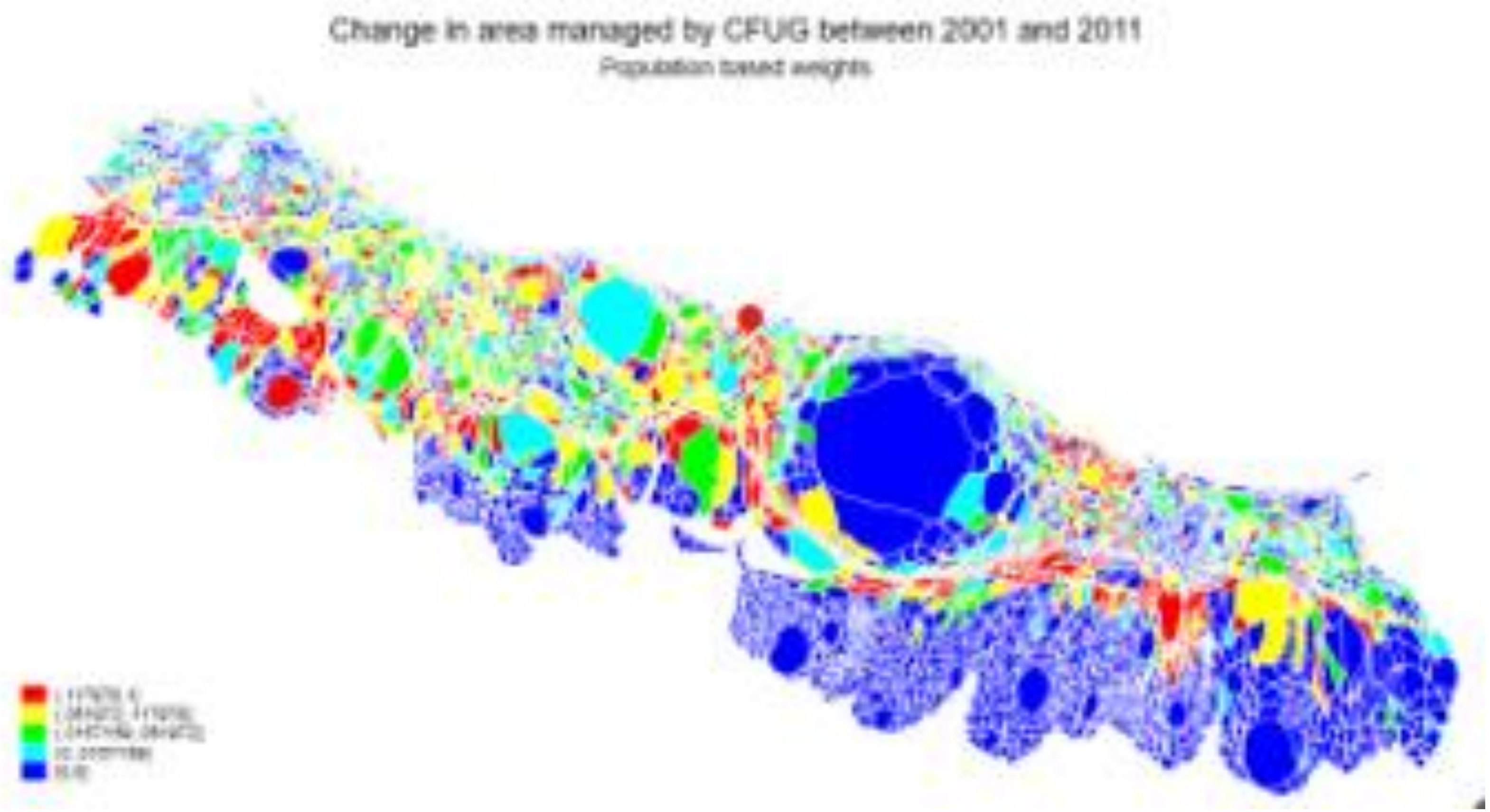
Change in Leaf Area Index between 2001 and 2011
Equal weight by administrative unit

Change in share of area managed by CFUG between 2001 and 2011
Equal weight by administrative unit

Here, one takes unit area as the unit of observation: each hectare or pixel has the same weight. We measure biomass per ha.

Here, one takes the village as the unit of observation. All villages have the same weight. We measure biomass per village.

And this is what one obtains when one uses population level observations each household has the same weight: we measure biomass per household



Change in area managed by CFUG between 2001 and 2011

Population based weights

Do not overestimate the real information you have:

- Suppose that you have information on vote shares per electoral circonscription that you want to correlate with circonscription characteristics (e.g. pro-Trump and share of evangelists).

- If you do not have individual data on vote, the level at which to place your analysis is the circonscription. You cannot make as if each vote in the circonscription was an ‘independent’ observation (not so rare a mistake in empirical papers).

- Be aware of the **ecological fallacy**: if you find a positive correlation between evangelists and pro-Trump, it simply means that areas with more evangelists are more likely to vote for Trump. It says nothing about individual behaviors.

- The ecological fallacy can sometimes lead to phenomena like **Simpson's Paradox**, where a trend observed in separate groups reverses when the groups are combined. (For example, when comparing two classes of a very different average scores, if a good score from the lower class joins the upper class, the average score in both classes may fall).

Why theory?

3.2 ‘Theory’ indicates how to measure your variables

Do not mix magnitudes that do not make sense. Make sure that your normalizations are consistent: is it in dollar? In X per dollar? In X per head? In %? In z-scores? In logs?… Measuring something implies that you already use an abstract concept, a ‘theory’.

Take a production function (physical capital, education and labour):

$$Y = K^{\alpha}E^{\beta}L^{1-\alpha-\beta}.$$

You can estimate this in log terms: $\log Y = \alpha\log K + \beta\log E + (1-\alpha-\beta)\log L$

or in dlog terms (growth rates!): $\Delta\log Y = \alpha\Delta\log K + \beta\Delta\log E + (1-\alpha-\beta)\Delta\log L$

or in per capita terms, dividing by L: $Y/L = (K/L)^{\alpha}(E/L)^{\beta}$

or the latter, in logs or dlogs,…

They are all consistent with the simple theory function that you have above. But many others are not or require another theory!!

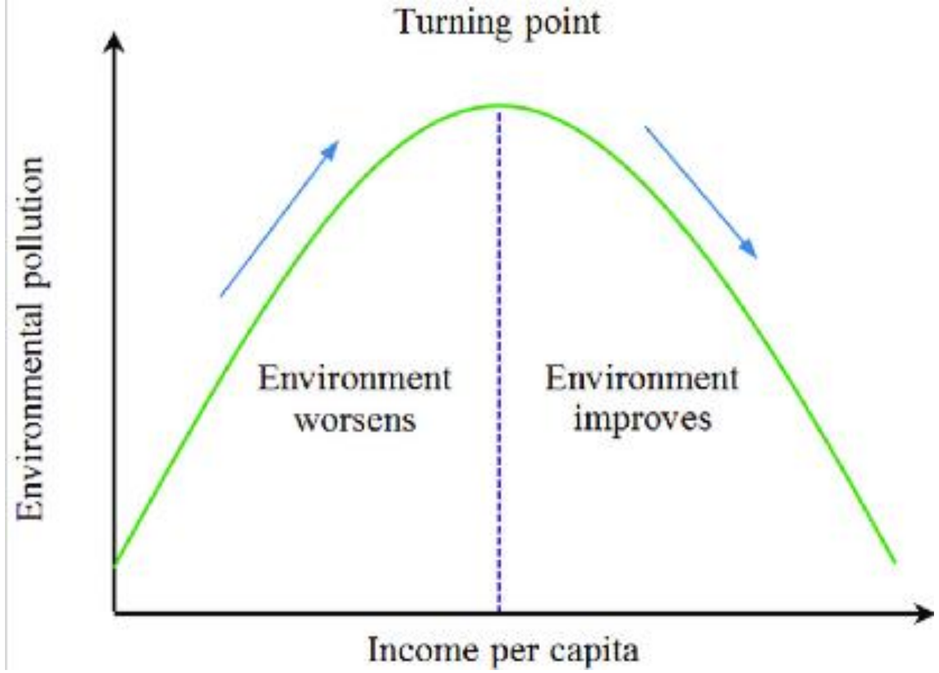
Why theory?

3.2 Theory indicates how to measure your variables

Take the EKC. One could hypothesize $P=\alpha Y+\beta Y^2$. P= total pollution; Y=GDP

- You can also test in per capita terms: $P/L=\alpha Y/L+\beta(Y/L)^2$ which implies a different structure (the L term does not simply divide the former equation),
- or in $ terms: $P/Y= \alpha Y+\beta Y^2$ which is yet another test!

Which one to choose is given by your theory, that you will have to defend!



- Or suppose that you find that the rate of change in forest (deforestation) is constant

- If you look at forest area, this implies that forest area follows an exponential process!

- Which approach to favor depends on what you really want to measure, ie your hypothesis.

- Take the issue of deforestation. Assume to start with that forest area depends on the population that needs place to cultivate.

- Using an accouting identity implies:
  - Forest area = A + B population,

  we want to estimate B, the land each person needs to cultivate
  - Or (Forest area/total area) = A + B population density
  - Or d(Forest area/total area) = B d(population density)
  - …

  But NOT
  - d(Forest area/total area) = B population density

  Which requires another theory (structural change? demographic transition?)…

1. Why theory?

4. To build a Testable Model

- Theory indicates also where to expect interactions between variables or **heterogenous effects.** Heterogeneity is guided by theory!

- Theory indicates at what level to use **fixed effects** (variation within instead of across) and clustering corrections. Think carefully about the variation within instead of across groups.

- Theory may inform about the **functional forms**. Note that Taylor expansions allow a linearization of more complex functional forms.

- Theory also indicates which robustness tests and **'placebo' tests** are appropriate for your question

- At the extreme, **structural econometrics** take the complex interaction between different behaviors seriously, and model them theoretically. But this is another game (see eg Foster, Rosenzweig, Munshi,…).

1. Why theory?

5. To interpret your Results

Empirical findings cannot be interpreted without theoretical grounding.

Mediation analysis describes the different paths of theoretical causality, that you measure empirically through direct and conditional correlations between variables (interesting but out of fashion).

Also, theory will guide you once you start exploring the **external validity** of your work. How can your results be generalized beyond the sample you just analyzed?

2. Choose the Right data

- **Find Quality Data Sources**: Use reliable data sources that are relevant, reliable, and ideally up-to-date. This includes identifying potential biases, missing values, or unusual observations that could skew results.
    - For instance, it is very hard to analyze top incomes or large landlords as they are rare by definition. If they are missing, this is a worry, but if in the sample, they may be over-represented, due to luck. Some questions require carefully chosen data sets or personal surveys.

- **Check for Sufficient Data**: Ensure your data has adequate sample size and enough **variation** (between and within clusters of information) in the phenomenon of interest to support statistically significant findings.

2. Choose the Right data

- **Understand the Data**: Explore the data before diving into analysis. Start with a good description of your data and the variables you are interested in.

  Be in particular careful with 'outliers' that tend to have a disproportionate influence in OLS (given the square of the distance). One solution to this is to start with censored data (drop the X% bottom and Y% top observations). Another, too often used, is to use the log transformation that reduces the influence of 'large' observations, but give a disproportionate weight to 'small' variations in small values: going from 0.001 to 0.002 has the same 'impact' as going from 100 to 200!! This transformation, unless given by theory, requires adequate justification and robustness checks. Moreover, the question of how to treat the 'zeroes' is there, people routinely use a transformation of the type log(x+1) to deal with that, but this is NOT a good practice.

- **Good graphs** are very useful, and sometimes sufficient (à la Piketty or à la Deaton). A good description is better than a bad model!
- Check the simple correlations that will support your enquiry.

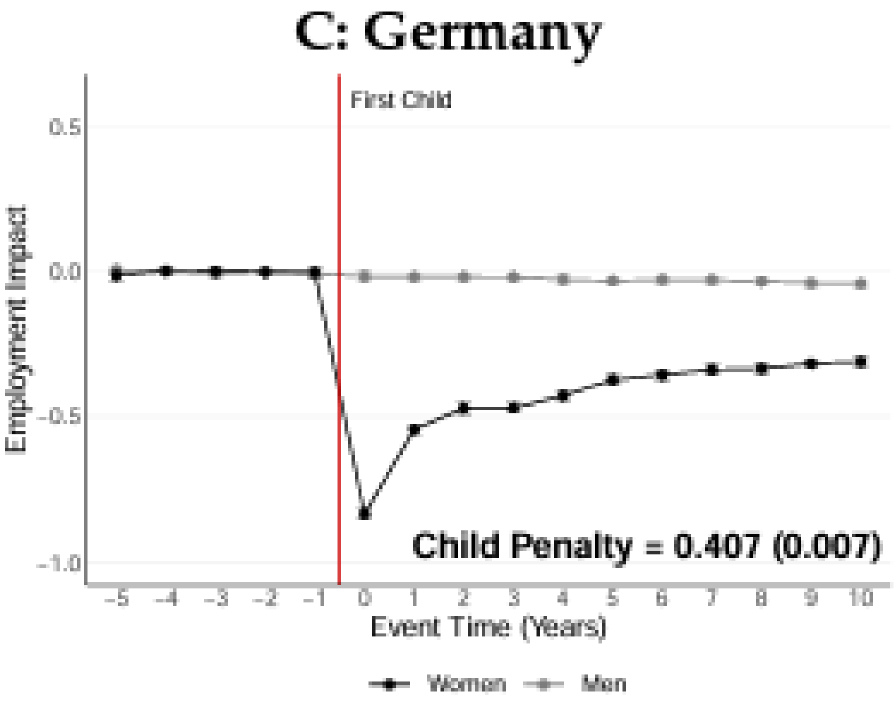
3. Build a Solid Identification Strategy… Or not

Do what you are thaught in your applied econometric classes:

- Address **Endogeneity**: Many economic relationships are endogenous, meaning they are influenced by unobserved factors or reverse causality. Choose methods that handle these issues, like instrumental variables (IV), difference-in-differences (DiD), or panel data models.
- **Find a Natural Experiment**: If feasible, look for a setting where a policy change or unexpected event acts like an experiment. This can help in isolating causality.
- For large data sets, also consider **non-parametric approaches**.

Or stay descriptive but do it well and make it interesting!

Kleven et al 2024



Employment rates of men (grey) and women in Germany before and after birth of the first child

Deaton 2008

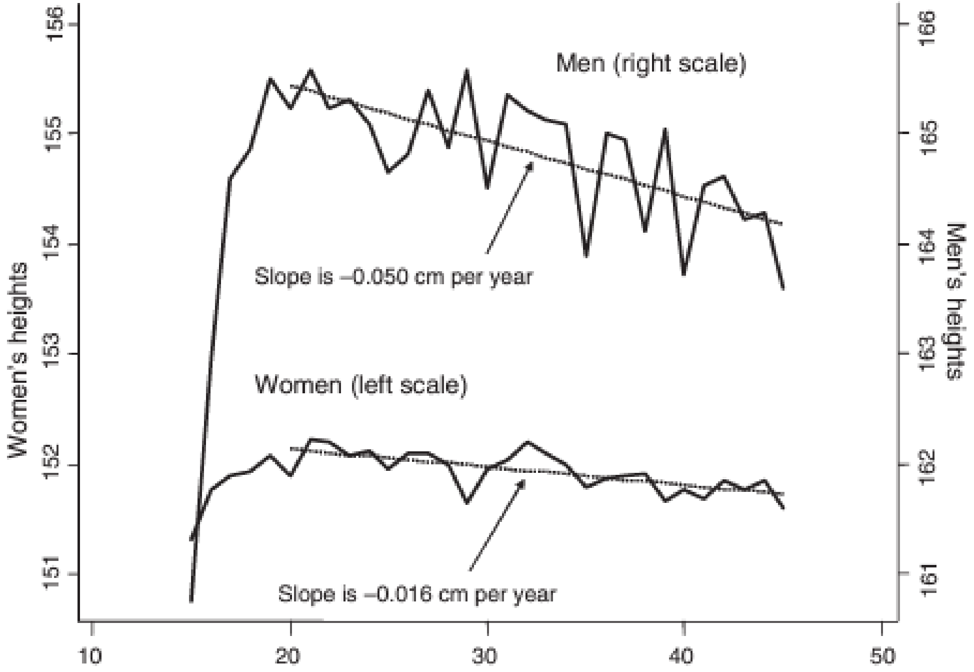


FIGURE 1. HEIGHTS AND AGE

Men’s and Women’s heights by age in India: is this a cohort effect (younger generation are taller) or a selection effect (taller people die earlier)?

Piketty 2011

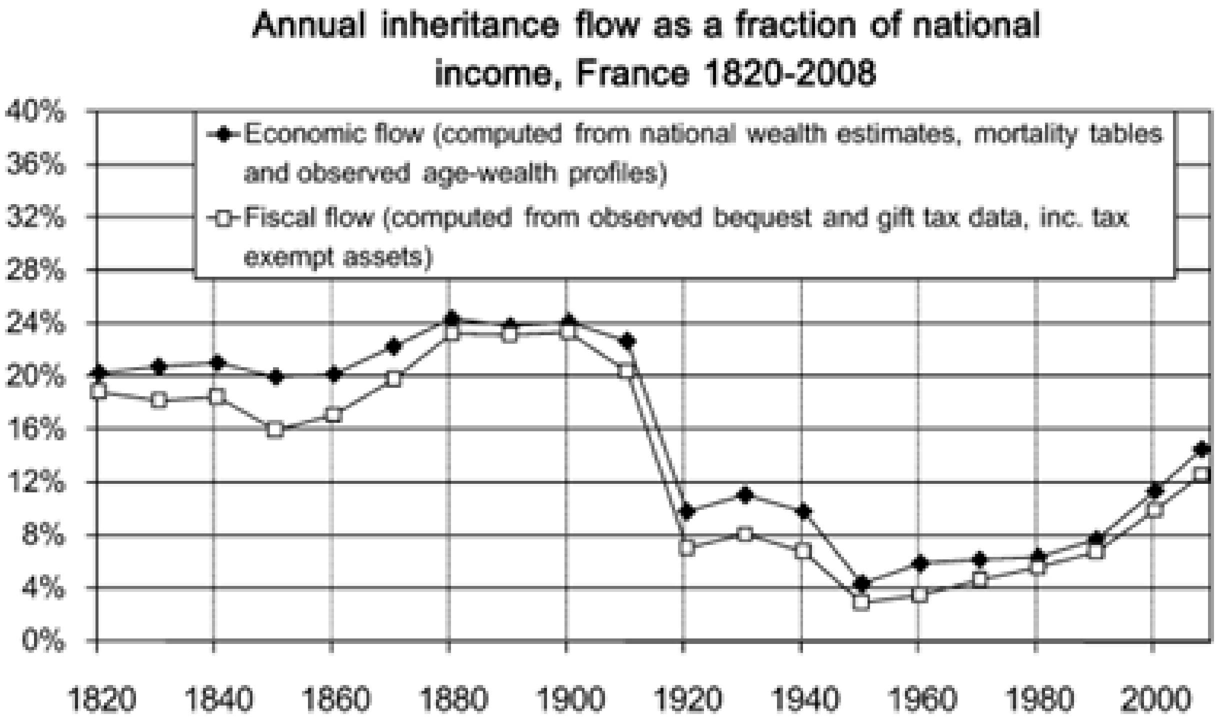


FIGURE I
Annual Inheritance Flow as a Fraction of National Income, France, 1820–2008

Anderson and Ray REStud 2010

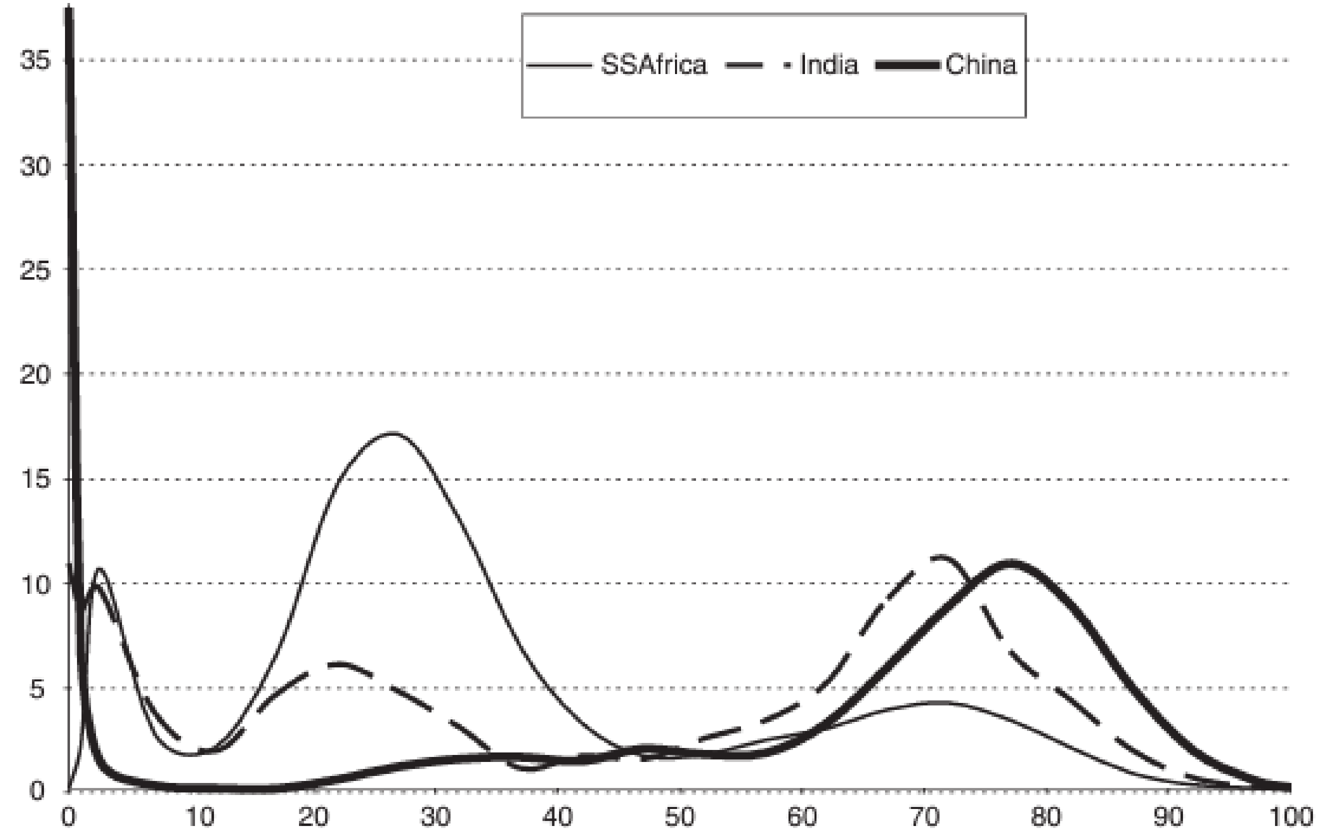


FIGURE 2

Missing women distributed by age (in %)

4. Check the model assumptions

- **Dont hesitate to question the ‘obvious’:** eg, is ethnic or caste identity a ‘given’, a real ‘exogenous’? Cassan (2015) questioned this and showed the fluidity in caste identity in India in the early 20th century (very interesting!).

- **Start simple:** do not overuse complicated techniques that are fashionable now. Start with simple description of the data, basic correlations and simple regressions, then choose the ‘right’ empirical approach to confirm (or not).

- **Check Model Assumptions**: Ensure the assumptions of your model hold in your data.
  - For IV models, the **exclusion restriction** should be carefully discussed and possibly investigated (both in theory and empirics). Reduced forms estimates should be produced.

5. Conduct Rigorous Data Analysis and Interpretation

- **Dont mess around** with tons of regressions to find the ‘best fit’! Always start with a proper simple model and estimate. Then think about your result and progressively improve on the model. In principle (but not always), robustness is a good sign and a healthy check!.
    - One of my co-authors always insisted: once you have designed your empirical ‘test’ (ie equation), try once and then stop!

5. Conduct Rigorous Data Analysis and Interpretation

- **Interpret Coefficients Carefully**: When reporting results, explain the **economic significance** of coefficients, not just statistical significance. For instance, a small but statistically significant effect may not be economically meaningful and is therefore of little interest. By contrast, with a large enough sample size, a null coefficient is also of interest!

- You can also use your estimations to build reasonable counterfactuals, using simple calculations, to assess the relevance or the importance of your results.

- If your results do not match your theory, or indicate a new avenue, go back to the theory, to understand it better (why is this so?) and probably design a new test. Empirics is also an iterative process in that sense: the data may surprise you.

6. Check for and Handle Potential Biases

- **Omitted Variable Bias**: Identify and include key control variables to avoid omitted variable bias, which can distort your results.

- **Selection Bias**: Be mindful of self-selection or non-random samples, and possibly consider methods to correct for selection bias, like Heckman correction models or propensity score matching. Or directly discussion the direction of those biases, and how they can affect the economic significance of your results.

- **Use Sensitivity Analysis and robustness checks**: Check how sensitive your results are to changes in sample, time period, or specification. For example, try excluding outliers, changing variable definitions or subsets of data , or using different estimation techniques.

7. Present and Report Results Transparently

- **Use Clear and Organized Tables**: Present tables and figures that make it easy for readers to follow your results. Use consistent variable names and provide detailed notes on interpretation.

- **Show Robustness and Limitations**: Report robustness checks and fully acknowledge and discuss the limitations of your study, including potential biases or untested assumptions. This adds credibility and helps future researchers build on your work.

- **Explain Practical Implications**: Discuss what your findings mean for policy or real-world applications. This helps readers understand the value of your research beyond academia.

8. Engage in Peer Feedback and Review

- **Seek Feedback**: Presenting your work to others can reveal new insights, potential errors, or alternative interpretations you might have overlooked.

- **Pre-Analysis Plans**: For more credibility, especially in studies prone to researcher bias, consider a pre-analysis plan, which specifies the hypothesis, methods, and analysis steps in advance. This is a bit of an ideal, as one often gets ‘surprises’ from the data, things that were unexpected and are interesting.

- **Consider Reproducibility**: Ensure that others can reproduce your findings by making data (if possible), code, and methodology clear and accessible.

9. Stay Updated with New Methods and Data

- **Follow Advances in Econometrics**: Methodologies in econometrics evolve, so keep up with developments like machine learning applications, causal inference techniques, and advanced panel methods that might add value to your work.

- **Utilize Software Efficiently**: Learn and use statistical software like Stata, R, or Python effectively to conduct analysis and create high-quality visualizations.

10. Write a Clear and Convincing Research Paper

- **Structure Effectively**: Use a logical structure, including an introduction, literature review, methodology, results, discussion, and conclusion.

- **Focus on Clarity**: Write in clear, accessible language, explaining technical terms where necessary. Aim to communicate complex results in a straightforward way (typically MA students is your favorite readership).

- **Summarize Key Takeaways**: Conclude with a summary of findings, policy implications, and potential areas for future research, giving readers a clear view of the study's contribution.

Part 3: Designing your own household survey

Household Surveys

It is an art, more than a science. Lot of learning by doing…

We focus on three issues here:

- Sample size and stratification
- Why and how to measure income?
- Questionnaire design

First principle: **BAD data = NO data**

Sample size

- Precision is NOT proportional to the size of the sample, but increases at the rate$\sqrt{n}$: To double precision, sample size has to be multiplied by four.

- Real trade-off: number of surveys versus the precision and care given to the questionnaires. Better spend more time on the questions and the enumeration: one imprecise answer on a relevant issue means one less observation.

- Typically, in most cases, 400 observations is already a lot.

Sample size

What if the phenomenon is not very frequent?

1) Increase the sample size to have ‘enough’ observations of the phenomenon you are interested in (100 is a reasonable target). But budget and manpower…

2) Switch to a case study, if the phenomenon is ‘rare’.

3) Oversample the object of interest. The rule here is to stratify on a **fixed characteristic**, not on the behavior itself.

Sample size

Example: Technology adoption

From field visits, we know the following:

+/- 20% big farms, 80% small farms

+/- 40% adopters in big farmers

+/- 10% adopters in small farmers

If 300 observations, we expect 24 big adopters and 24 small adopters, which is a bit small. What can be done?

Sample size

The initial distribution is:

	Big	Small	Total
Adopter	.08	.08	.16
Non-adopter	.12	.72	.84
Total	.20	.80	1.00

Sample size

If stratifying to oversample the adopters with half adopters, half non-adopters :

	Big	Small	Total
Adopter	.25	.25	.50
Non-adopter	.07	.43	.50
Total	.32	.68	1.00

What is the probability of Adoption if Small?

Sample size

What is the probability of Adoption if Small?

P(Adopt|Small) = 0.25/0.68 = 0.38

and P(Adopt|Big) = 0.25/0.32 = 0.78

Which are both seriously WRONG! The true probabilities are 0.10 and 0.40.

Fitting the regression line:

$$\text{Adopt} = \alpha + \beta \text{ Big}$$

β = 0.40 (instead of 0.30)

as β is simply the difference in the conditional means.

Sample size

- Lesson: you must over-sample on a fixed characteristics, not on the behaviour. Many papers do that mistake, and it IS a mistake!

- In this case, over-sample the big farmers, for instance 50% big, and 50% small farmers. With the same sample size fo 300, you expect 60 big adopters and 15 small adopters, and conditional probabilities are right.

- Alternatively, if prior knowledge of the proportions of farmers, and the proportion of adopters, one can reweight the probabilities and avoid biases. The problem comes from the fact that these are usually unknown!

Stratification

- There are some statistical formulae that can help, but they usually require prior information on the variations in the variables you are interested in.

- Why stratify?

- Oversample a rare behavior

- A really random survey is too expensive and in general infeasible. Typically cluster by villages, schools,…

- If the phenomenon varies given exogenous characteristics (urban/rural, geography, population,…), you may use a non representative sample that ensures enough variations along those dimensions (eg, forestry in India).

Stratification

When using clusters, the real cost is the travel cost to the clusters. How many clusters are desirable? The **source of variation** is crucial:

- If most of the variation is between households within villages, then few clusters with a lot of households.

- If most of the variation is between villages, with households in the same village behaving in the same way, then many clusters with few households per cluster.

Statistical corrections are important. If interest is in variation in behavior within the cluster, cluster fixed effects should be used. But not always! (If the number of clusters is large enough (50), cluster corrections should be used.)

Measuring income

- Income = ‘the’ measure we are typically interested in

- Three approaches:

  - Consumption
  - Income
  - Permanent income

Measuring income

- The 3 measures are different!
  - **Current income** responds well to short run shocks or incentives.
  - **Expenditures** (particularly food) are more stable across periods.
  - **Permanent income** is more a measure of long run wealth.
- The appropriate measure depends on the question to be addressed.

Measuring income

Consumption expenditures

- You have to list all items in detail (e.g. different types of rice), then report all that has been consumed and purchased (price and quantities) over a given period.
- Example…

SECTION 7. CONSUMPTION EXPENDITURE PART A RECURRENT EXPENSES AND HOME PRODUCTION

				FOOD PURCHASED					HOME PRODUCTION				IN KIND
1. Have you consumed …(Food)… during the past month? PUT A CHECK (✓)IN THE APPROPRIATE BOX FOR EACH FOOD ITEM . IF THE Q. 1 IS YES, ASK Q. 2-8.				2. Over the past month, did you purchase …(Food)… ? *IF NONE LEAVE BLANK AND GO TO 5*	3. How much did you purchase?		4. How much did you spend in total to buy this quantity? RUPEES		5. Over the past month, did you consume …(FOOD)… that you grew or produced yourself? IF NONE LEAVE BLANK AND GO TO 8	6. How much did your household consume of …(FOOD)…?	7. How much would your household have to spend in the market to buy this quantity of …(FOOD)…? RUPEES		8. What is the total value of the …(FOOD)… consumed that you received in –kind over the past month (wages for work, gift,...) ? RUPEES
	NO	YES			QUANTITY	PRICE PER UNIT				QUANTITY			
1.GRAIN/CEREAL													
Rice PDS													
Rice other sources													
Chira													

Measuring income

4 problems:

- Based on recall. Can sometimes be solved by a daily expenditure booklet.
- Measuring self-produced commodities: Which prices to use? In the absence of local markets and local prices, use shadow values?
- Measuring quantities properly: particularly for self produced items or purchased in ‘large’ quantities: amount of oil per week? Of salt? Of firewood?
- Very long and Very boring for the respondent.

Measuring income

Measuring income directly.

- Fine in a urban environment with regular wage earners or employees.
- For self-employment or irregular jobs, this is much harder.
- With self-employment, there is confusion between values of production, sales and income. The standard way is first to estimate the value of production:

SECTION 2 INCOME /EXPENDITURE **PART A INCOME FROM AGRICULTURE (self-employed)**

1.What is the total amount of land cultivated last year : □ 2. Unit used : □ Khatta…1 Bigha…2 Acre…3 Hectare…4 Other…5 specify :____________

CROP CODE	3. Area under crop Bighas.	4. Unit used Khatta..1 Bigha…2 Acre…3 Ha…4 Other…5 specify :_________	5. Quantity Harvested *(1 quintal =100kgs)* in kgs	6 Home Consumpton kg per year	7. Quantity Sold *If not sold then '0'.* (kgs.)	8. Price per unit (price received) *If not sold then 'NA' If given in kind, price at which it could have been sold on the market* (Rs./kg)	9. Marketing mechanism. *If not sold then 'NA'.* Within the village…1 Local trader…2 Local market…3 Wholesale market elsewheere…4 Government/BDO…5 Other…6 Specify:

CROP CODES

Cereals 1. Maize 2. Wheat 3. Paddy 4. Barley 5. siur/marsha/chalai 6. Phoolan 7. Ogla 8. Phapra 9. kodra/madua 10. Gangdi

Pulses 11. rajma 12. mash
13. kulth 14. soyabean 15. masoor

Vegetables 21. potatoes 22. peas 23. beans 24. cabbage 25. tomatoes 26. garlic 27. katcha aloo 28. chillies 29. onion

…

Measuring income

- Then, list all the costs (rents, fertilizers, equipment hiring,…). Be careful that the time units used are compatible (year, month, harvest,…)

- For crops, one standard procedure to avoid recall problems is to visit each plot in the fields, and carry out a plot-wise survey.

- Then, using accounting rules, reconstruct the ‘income’ of the farm…

- We face similar difficulties as with consumption expenditures.

Measuring income

Measuring permanent income

- Collect information on wealth.
- Major assets, like land and house
- Durables
- …

Problem: how to aggregate into a meaningful measure of wealth?

- Principal Component Analysis: problem of arbitrary weights?
- Accounting method: use or guess prices?

Measuring income

- With all measures, two other problems arise:

  - Under-reporting bias: respondents forget expenses, crops,… and therefore under-report. Problem arises if this bias is correlated with some respondent characteristics that matters (IQ, wealth,…) (see below)

  - Seasonality: matters a lot in rural areas, so the appropriate timing has to be thought over.

Measuring income

- Errors do not matter that much in a ‘large’ sample to measure ‘averages’ over a subset.

- Also fine if income is just used as a ‘control’ in the regression (in which case a quick measure is preferable).

- This is NOT fine if you intend to measure the impact of income on a measure of interest, or the impact of an exogenous change on income and consumption (‘welfare’).

Measuring income

- Example: suppose that a change in the environment (e.g. microcredit) increases consumption by 3%. The measures above have errors of at least 30%…

- Except in a very large sample, the effects cannot be estimated.

- Another worry is that biases in income reporting may themselves be correlated with respondent's characteristics that matter (IQ, vigilance, self-confidence,…) for your results! Asking for too much detail may only exacerbate those biases.

Measuring income

- The time spent on measuring income in the interview is better used for more interesting issues.

- Either use a rough measure of income (no income brackets, please!)

- Or a good proxy for income and wealth: land size, occupation, number of meals a day,… These proxies can be defined during the presurvey: they should make sense, and vary enough across individuals.

Measuring income

- Or measure correctly **food expenditures**:
  - measure adequately ‘permanent’ income,
  - less sensitive to transient shocks,
  - Provide adequate ordering of households in terms of welfare/income

- Note:
  - You may even concentrate on cereals only
  - Aggregate through prices, or caloric intakes (used in poverty measurement).

Questionnaire Design

Golden rules:

- An interview is a **dialogue**, a conversation.
- A good survey needs a very good **pre-survey**

Questionnaire Design : the questions

Ask only **necessary** questions and stress facts, not opinions.

- Express your research questions
- Think about the different hypotheses
- List the information you need to test these
- Define the variables that give this information.
- Find the best question to ask to measure that variable
- Think about data entry

But, at little marginal cost, can you address other questions? If yes, do it.

Questionnaire Design : the questions

Don’t ask stupid questions, to which you cannot yourself give a proper answer!

Check:

Can you answer that question: ‘what is your income per month?’ ‘how much food did you consume last year?’,…

Use words that make sense:

Example: an interest rate is a complicated concept. What is the interest rate if for 100 Rs borrowed, you reimburse 8 Rs a month for 5 years?

Questionnaire Design: the questions

Don’t ask embarrassing questions on delicate topics: land conflicts, maternal history, contraception, domestic violence,… if they are not necessary.

If necessary:

- Find the best way to get the information: selected village informants, separate interviews with the members of the household, use of female enumerators,…
- Justify these questions to the respondent for your research or policy objective.
- Avoid hit-and-run approaches, where delicate questions come at the end of the questionnaire: this shows a lack of respect.

Questionnaire Design : the questions

Questions are self contained, complete and non-ambiguous.

A) Specify the who, when, where, over what period, units of measurement…

Example: ‘How many children do you have?’

Does it include: dead children? Older ones who left? Children fostered in? Children from other spouses?…

Questionnaire Design : the questions

B) Define clearly at the start the unit of analysis. For a household, do you include:

- the son who lives next door in a separate house but with the same courtyard?
- the husband who works in town for 10 months a year?
- the eldest son who is in boarding school and returns only for holidays?
-…

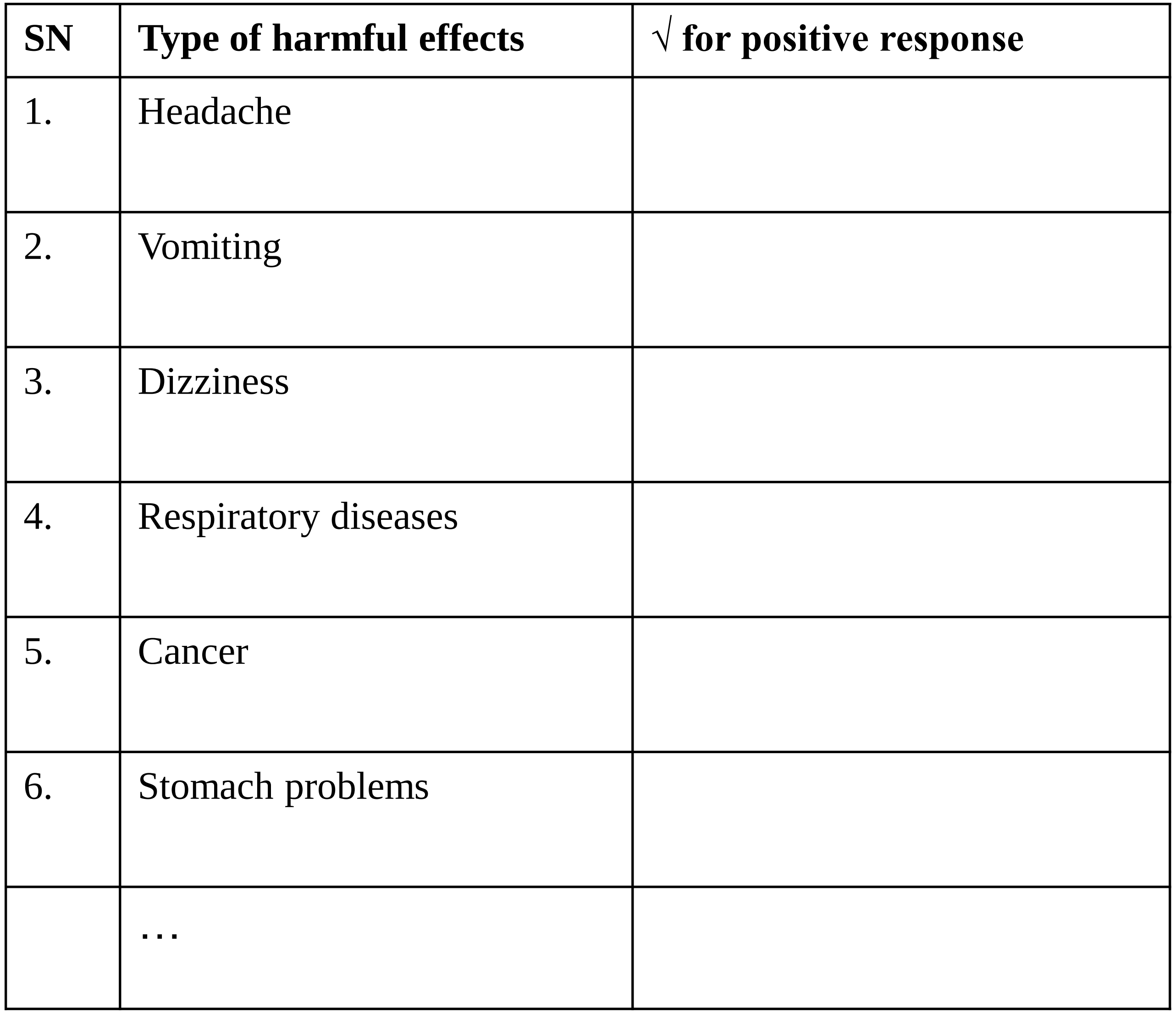
Questionnaire Design : the questions

C) Have question with a unique meaning.

- Useful to ask enumerators how they understand each question,
- and how they translate it best in local language. The Questionnaire is written in local language.
- These things can be sorted out while training the local enumerators, and discussing each question with them.

D) Avoid loaded questions, be as neutral and factual as possible. Be aware of the framing effects, even with ‘neutral’ questions.

Could you please share any of the following harmful health effects of applying inorganic chemicals on you and your family members in last year?



SN	**Type of harmful effects**	**√ for positive response**
1.	Headache	
2.	Vomiting	
3.	Dizziness	
4.	Respiratory diseases	
5.	Cancer	
6.	Stomach problems	
	…	

A bad example…

Frames: Example: (Kahneman and Tversky)

Problem 1 (Save): *Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:*

✓ *If Program A is adopted, 200 people will be saved.*
✓ *If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.*

Which of the two programs would you favor?

_72%____ ___28%____

Program A *(Risk Averse)* Program B *(Risk Seeking)*

Frames: Example: (Kahneman and Tversky)

Problem 2 (Die): *Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:*

✓ *If Program A is adopted, 400 people will die.*
✓ *If Program B is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die.*

Which of the two programs would you favor?

_*22%*____ ___*78%*____

Program A *(Risk Averse)* Program B *(Risk Seeking)*

Questionnaire design: the answers

Avoid open questions, use precoded closed ones!

- Open questions = badly done questionnaire or bad pre-surveys.
- If opinions are needed, be very careful about the precise phrasing and possible answers. The pre-survey should focus on those questions.

Keep as much as you can consistent time, value and quantity units for similar issues (weekly for frequent consumption, last month for work, last 12 months for agricultural production,…).

Questionnaire design : the answers

Pre-code everything answer you can code. Codes avoid mistakes, misinterpretations…

- Avoid unequally detailed answers, which forces the use of the coarsest category.

- Do not use coarse codes when finer ones are better understood: for age or education, use the number of years and not intervals 1-4, 5-8,…

- Always leave a ‘other specify’ category.

- Do not leave blanks in a questionnaire, unless you allow skipping questions: ‘If no, go to section Z.w.24’. Otherwize, use a separate code: NA, 999,…

Questionnaire design : the answers

With prepared answers, there is a choice between:

a) the respondent gives an answer, that is then matched with the pre-coded ones (best practice).

b) enumerator reads all possible answers:

- useful for difficult questions
- respondent may jump on the ‘first answer’. Make sure that all answers are read.
- Ideally, randomize the order of the answers across questionnaires.

Have answers that are closest to the actual life of the respondent: if a laborer is paid in Kgs of rice, ask how many Kgs, not the value in Rs.

Use other fonts for the answers and for the remarks that the enumerator should not read.

8. What is the price per unit that you received for …? (Rs./kg) *If not sold then ‘NA’* *If given in kind, price at which it could have been sold on the market*	**9.Where did you sell …?** *If not sold then ‘NA’.* Within the village…1 Local trader…2 Local market…3 Wholesale market elsewheere…4 Government/BDO…5 Other…6 Specify:_________________

A bad example…

- **VILLAGE INFORMATION**
- 1. Name of the village
- 8. Type of approach road
- 9. Distance from nearest town
- 10. General description of topography
- 11.Modes of transport used by villagers
- 12. Area of village
- 16. Population
- 17. Sex ratio
- 18. Number of households
- 19. Average landholding
- 20. Type of houses (walls and roof)
- 21. Main occupation/s, main crops grown
- …

Questionnaire Design : the answers

If you want opinions:

- be very careful about the precise phrasing. The pre-survey should focus on those questions.
- list all possible answers.

Open questions are very useful to get interesting stories and background information. They require proper training and, sometimes, a separate enumerator.

Questionnaire Design : the answers

Check answers when feasible:

- Can you read and write (yes/no)? Can be replaced by asking the respondent to recognize bank notes/write two words on a piece of paper/ signe the interview…
- For agricultural surveys, check the cultivated plots (size, crop patterns, trees,…)

Questionnaire Design: the structure

Be Logical!

- The questions are logically related, and follow a conversation mode. Don’t jump from one topic to another!
- Pre-testing of the questionnaire is very useful to re-organize the sections.

Questionnaire Design : the structure

Be consistent:

- Keep the same codes (eg, for Yes/No), tables, IDs, presentation, listings,… throughout.
- If needed on a paper questionnaire, have a grouped list of codes at the end of the questionnaires, if they are too long, or are repeated across the questionnaire.
- For each household member, have a list with an ID code that you keep throughout. You may register them on a separate page for recall of the name/gender/age, or precode them on the tablet
- Never change your questions during the survey! Add a new question if you realize that a key info is missing.

Questionnaire Design : the structure

Beauty matters!

Be esthetic, practical and concise!

- Lay out should be clear and attractive
- Tables should be used for a synthetic view of the answers
- Keep enough space for each answer…

1.2 Information about household head

A. Religion				**B . Ethnicity**				**D. Were you a pre-tsunami boat owner**		
Buddhist	1			Sinhalese	1			Yes	1	
Hindu	2			Sri Lankan Tamil	2			No	2	
Muslim	3			Indian Tamil	3					
Christians	4			Sri Lankan Moor	4			**E. If Yes , mention the boat type**		
Roman Catholic/ Other	5			Malay	5			Offshore Multiday	1	
				Burgher	6			Day boat with Inboard Engine	2	
				Other	7			FRP boat with Out Board Motor	3	
								Traditional Motorized Boat	4	
C. Has household head changed after tsunami?				Yes	1			Traditional Non Motorized Boat	5	
				No	2			Beach Seine Boat	6	

Questionnaire Design: the respondent

Ask the relevant person:

Examples:

Child vaccination and illnesses to the mother,

Secondary school: to the child if possible,

Land tilling technique: to the male head,

…

Think also about the best level at which a question should be asked: group, village, husband, individual, neighbour,…

Questionnaire Design : the respondent

Think about your respondent! He gives you time and effort:

- The questions should be friendly and empathic.
- The questionnaire should not exceed 90 minutes.
- Your enumerators should be trained to explain why the survey is important.
- and ensure anonymity the best way you can (sometimes it is better to say nothing).

You may wish to give a gift: bucket, soap, scarf, bag of rice, money… Be careful about the appropriate practice and th fact that not all households in the village will be interviewed.

Questionnaire Design : the respondent

Some questions are used to help him answer or remember.

Detailed questions are useful here if they help. There is a trade-off:

- Go field by field to get information on crops and production.
- Go through the names and age of the children to make sure that you have the right number of children.
- Go through the two harvest seasons for information on labour hiring.
- …

Data entry

In most cases, you will use tablets with a coded questionnaire.

If you use an paper questionnaire,

1. Think about the logistics involved (who collect the questionnaires, how to transport them,…)
2. Think about having separate pages with all the codes, and possibly the household roster (as a check) on separate sheets
3. Data entry is an issue.
   - Do it on the spot, as much as you can: allows you to discuss with the enumerator to clarify imprecise answers, or send him back.
   - Entry teams of two can be used: a reader and a typer. For large surveys, a possibility is double entry with cross-checking, but this is costly.
   - Make sure to have proper cluster/household/individual identifiers (identity numbers!). This is essential…

Questionnaire Design: some last thoughts

Do not start from scratch!

- Use World Bank LSMS: http://iresearch.worldbank.org/lsms/lsmssurveyFinder.htm
- Check two useful books:
  - Paul Glewwe ‘Designing household surveys’, WB, 2000.
  - Angus Deaton ‘The Analysis of Household Surveys : A Microeconomic Approach to Development Policy’, WB, 1997.

Keep an eye on your enumerators!

- Check the questionnaires they fill in in the field, discuss stories, check inconsistencies, send them back if needed (be tough),
- Pay them per day or per questionnaire?
- Supervise on the field! Do not delegate.

Do your pre-survey carefully:

- Do not impose impossible deadlines, and allow enough time for the pre-surveys (typically 3 to 4 months with gradual improvements in the questionnaire) and the initial field visits 5 or 6 months earlier to ‘explore’ an idea.
- Presurvey usually suggest that the questionnaire is too long: with some practice, the length of interview is reduced to about a half.
- The pre-survey is essential: correct phrasing of the questions, correct units, correct answers and codes,…

Questionnaire Design

Golden rules:

- An interview is a dialogue, a conversation.
- A good survey needs a very good pre-survey