Summer School onCapability and Multidimensional Poverty
28 August-9 September, 2008
New Delhi, India

HDCP-IRC
The Human Development, Capability and International Research Centre
Instituto Universitario di Studi Superiori
www.iusspavia.it

OPHI
Oxford Poverty & Human Development Initiative
University of Oxford
www.ophi.org.uk
Overview of empirical analysis: primary and secondary data

Enrica Chiappero-Martinetti
University of Pavia and HDC-IRC, IUSS
29 August, 2008
Contents

• to sketch a general taxonomy of data requirements/data availability for implementing empirical analysis based on CA
• To discuss pros and cons of alternative strategies, namely:
  – Primary analysis
  – Secondary analysis:
    • Macro data
    • Micro data
• To list some of most commonly applied statistical techniques
• To briefly review the main empirical papers based on CA
Overall goals

• to disseminate data sources and tools for stimulating new research and improving quantity and quality of empirical work in this field
• to collect and post on the HD-CA website a selection of data sources and tools (e.g. questionnaires, software) useful for developing empirical analysis
• to share some basic information on each of these sources and tools so to facilitate the choice among alternative options
• To invite students and researchers in the HDCA network to help to integrate, extend and update this database
What the researcher potentially needs?

- Plurality of evaluative spaces: agency-empowerment, standard of living - well-being, capability, achieved functionings
- Plurality of domains (health, education, housing, employment, participation etc.) and multiplicity of variables for each domain (qual/quant, obj/subj)
- Plurality of units of analysis and human diversity: individual, family/household, specific sub-groups of population (i.e. women, children, elderly, by caste, by ethnic or religious group), whole community
- Plurality of contexts: socio-economics, geographical, cultural
Empirical literature based on the CA

(see table 1)

• Take into consideration a plurality of domains (from 3 to 14)
• Pay attention to human diversity, in particular to gender
• Are equally addressed to developing and developed countries
• Are primarily focused on achievements (“second-best choice”? even if some attempt to measure capabilities has been done
• Make equally use of both aggregate figures and micro-datasets
• Are generally based on secondary data with some interesting experience of gathering primary data
What available data sources potentially offer

**PRIMARY ANALYSIS**

(see table 2)

Generation of new dataset in order to address a specific research question (e.g. measuring capability, agency or empowerment, etc.) using direct methods (interviews) or indirect methods (observations, focus groups, etc.)
Pros →
1. data tailored to the specific research question
2. more in-depth analysis focusing on specific contexts/situations/individuals
3. more attention to non-metric, qualitative variables
4. respondent can play a more “active” role (i.e. selection of relevant functionings)
Cons →
1. costly and time-consuming
2. validity and reliability of the analysis difficult to verify
3. small sample size, statistical significance and disaggregation
4. comparability across time or across space
5. choice of the statistical techniques affected by the sample size and the nature of the survey (qual/quant data)
SECONDARY ANALYSIS

Based on an existing dataset and aimed to address a research question (e.g. measuring functioning or capability “proxies”) that is distinct from that for which the dataset was originally collected.
Pros →
1. generally, large-scale surveys, statistically representative, 
   disaggregate investigations
2. it helps researcher to focus on a broader concern
3. availability of multiple sources can allow comparison of trends over 
   time and complementary data analysis
4. the same dataset can be analysed from different disciplinary 
   perspectives contributing to multidisciplinary understanding
5. many datasets are freely available on the web
6. can be used for a preliminary investigation for a primary analysis
Cons →
1. to tailor the research question to fit the data
2. large datasets can involve a great amount of statistical analysis and techniques
3. secondary datasets largely differ in term of size, depth of interviewing, and complexity of the topic, content, degree of geographical detail, time-specificity
Major sources of secondary analysis in social sciences

1. **Macro (aggregate) data**

(e.g. population census, large continuous, regular and official surveys, datasets derived from administrative records such as marriage registers or tax records), most available in terms of published reports on methods used and main results. Statistics or tables are sometimes produced from micro-data relating to broad groups, areas or categories

(see table 3.1)
Pro →
Usually harmonized datasets (e.g. UN, OECD, EUROSTAT statistics)
both time-series and across-countries that allow international and
inter-temporal comparisons

Cons →
Aggregate analysis can hide deep inequalities and internal disparities
among subgroups of population and/or individuals
2. **Micro-data**

(multipurpose surveys, longitudinal or ad-hoc datasets with a fairly specific focus): they contain information on individual respondents to a specific inquiry

(see table 3.2)
Pro →
data gathered through sample surveys are generally more informative and allow to make more refined analysis

Cons →
they are more complex and “time-expensive” at computational level
Briefing on statistical techniques:
1) Scaling and ranking solutions

(e.g. HDIs) normalization of indicators (e.g. life expectancy, literate rate): usually, linear function with top and bottom limits (e.g. 1 and 0); aggregation: generally, arithmetic or weighted average. Nb: it assumes perfect substitutability between the functionings.

**Strengths**
- It allows ranking countries or regions and assessing their performance in accordance to each dimension or to the synthetic index.
Weaknesses:

- not a statistical solution but a tool largely based on analytical considerations
- difficulties in defining bottom and top limits for each dimension; disagreements on weighting (arbitrariness and assumption on the relationship between dimensions)
- it is not as successful in providing a solution for interpersonal or interhousehold comparisons
- It requires intervals and ratio measures, not able to deal with nominal variables or ordinal scale, particularly common at a micro level.
2) Fuzzy set theory

Extension of classical set theory or crisp method; it provides a mathematical framework for handling categories that allows partial membership (or membership in degree) usually (but not necessarily) in a range between 0 and 1.

**Strengths**
- it is able to preserve complexity and vagueness of categorical concepts and handle well qualitative, quantitative variables and linguistic attributes
• it allows to combine indicators by a large set of aggregation operators (union, intersection, average operators)
• It can be combined with other statistical techniques, by making statistical inferences and models testing

Weaknesses:
• It is also a “non-statistical solution”, requiring significant analytical considerations and a broad knowledge of indicators and context in order to define the membership function in an appropriate manner
3) PCA & FA

Functionings (or overall well-being) as latent variables (i.e. combination of a range of indicators using either the correlation matrix (FA) or the eigenvalues (PCA). Can be used as exploratory or as confirmatory analyses as well as reduction technique

Strengths

• less affected by the researcher’s arbitrary decisions, being the expression of the statistic relation between the indicators (weights are derived from variance of the indicators themselves)
• the final factors score can be used as a continuous variable in other statistic analysis.
Weaknesses:

- mainly data reduction techniques to be used for the selection/identification of functioning more than for representing and aggregating variables and well-being dimensions
- They implies perfect substitutability.
- The final factors score tends to be difficult to interpret in factor analysis, requiring the use of the same scale of measure in each indicator.
- The selection of the indicators remains a key issue requiring significant analytical decisions.
• Variables should be quantitative at the interval or ratio level: ordinal scale variables need to be interpreted in cardinal sense, although their mean or variance do not have real meaning (but categorical variables are not suitable for factor analysis).

• The basic assumption is that data have a multivariate normal distribution (not necessarily true) and observations should be independent.
4) Clustering analysis

It explores data in order to find clusters of variables that represent the same underline functioning (nb extension of exploratory factor analysis since it uses the statistic information contained in the entire distribution and not only the covariance or correlation matrices of the data).

Strengths

• useful non-parametric tool for identifying attribute groups to be used in multidimensional welfare analyses’
• it may reduce the chance of double counting highly similar attributes
Weaknesses:

- The unit of measure can be a problem when using ordinal indicators, in which case they would have to be given a cardinal interpretation.
- Similarly to FA, the result depends on the theoretical relevance of the selected attribute.
5) Regression Approach

- 5.1) To explain the functioning achievement or overall well-being, by income and socio demographic characteristics or contextual factors (conversion factors)
- 5.2) To estimate the well-being production process (conversion rates)

**Strengths**

- It allows studying how functionings achievements are related to resources and conversion factors
• The correlation between functioning measure and income, controlled for other variables (e.g. Socio-demographic background) allows proving the complementary value of welfare analysis in the functioning space.

**Weaknesses:**

• Difficulty to use ordinal variables (only possible with dummy variables).

• Unidirectional definition of causal relations which is quite problematic in the capability approach where functionings are interrelated.
6) Structural Equation Models

They allow estimating a series of separate multiple regression equations simultaneously. Two steps:

i) The model uses several variables for a single independent or dependent latent variable (similarly to FA).

ii) causality is considered in a more complex way, assuming dependence relationships simultaneously (important where a variable can be a cause of a particular variable as well as a consequence of another)
Strengths

• It combines factor analysis and regression analysis in a single step
• It develops a more systematic and holistic view of problems, where causality is seen in a more complex way assuming dependence relationships simultaneously.
• It deals with measurement error in both endogenous and exogenous variables
• Ordinal variables can be considered by transforming them into latent continuous variables
Weaknesses

- the results are difficult to interpret for policy matters
- similarly to FA, the selection of the indicators needs to be well grounded theoretically
References and useful information sources

1. Section in the HDCA website promoted by the QRM-TG:
   
   [http://www.capabilityapproach.com/empirical.php?sid=fe1948253a9d78edd0cae3f22875de17](http://www.capabilityapproach.com/empirical.php?sid=fe1948253a9d78edd0cae3f22875de17)


3. ECM and J.M. Roche forthcoming paper
   “Operationalization of the capability approach, from theory to practice: a review of techniques and empirical applications”

4. JHD section on most recent papers and books (nb: not only in English)
Student exercise

With reference to your specific field of research or interest:

1. Are both primary and secondary analysis potentially equal feasible options for you? One is better than the other or simply you don’t have any option?

2. Identify what are the main pros and cons of both options (primary vs secondary analysis) for your specific research project.

3. If you already did some empirical analysis, how have you choose between primary and secondary methods and, in the latter option, how have you select your data among available datasets?