Describing data

LECTURE 15

C4D2 TRAINING 1

1

The dissemination phase

- Last step of the journey
- Typically, after survey and data processing are concluded, a report is released, aimed at describing main findings from survey

C4D2 TRAINING 2

2

Dissemination of what?

Topics vary according to contents of the survey, target audience, etc., but there are some common elements:

1. Background information on sampling
2. Descriptive statistics (roughly corresponding to survey modules)
3. In the case of income and expenditure surveys, measures of inequality and poverty

Next slides cover these 3 points, with tips for effective presentation and examples.

C4D2 TRAINING 3

3

A note on the target audience

- Addressing a technical or academic audience is different than informing laypeople or the media
- NSOs often use different communication instruments for these different audiences, e.g. technical reports vs. press releases
- But 'technical' is not a synonym for 'obscure'
- This lecture covers general principles that are useful for different audiences

C4D2 Training 4

4

1. Background information

C4D2 Training 5

5

The need for background information

- Survey reports are usually designed to be accessible to a non-technical audience
- But technical background information must still be present, to inform more advanced readers and facilitate comparisons over time and across countries
- Background information can be presented separately from "core" results, as introductory chapters, appendices, or even a companion document, but must not be omitted

C4D2 Training 6

6

What not to miss

- Reports should document at least the following survey design features and processing choices:

a. Sampling design
   Sample size, stratification, representativeness…

b. Data collection and processing
   Fieldwork, outlier detection and treatment, data imputation…

c. Definitions of economic concepts and aggregates used
   E.g. disposable income, total household consumption, imputed rent… May be presented as a glossary

7

Documentation on sampling design

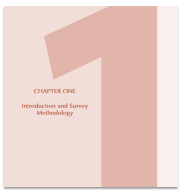
what to include

- Sampling design report with
  - Allocation of sample into strata and indication of excluded strata, if any
  - Estimation formulas (selection probabilities and weights)
- Household listings forms
- Sample frames
  - For the first sampling stage/s: list of all sampling units
  - For the last sampling stage: list of all households in each sample point
- Non-response rates
- On the survey datasets
  - Sampling weights

8

Kenya, 2015

Kenya Integrated Household Budget Survey (KIHBS)



Parameters	2015/16 KIHBS
Sample design	
Survey Domains	National, 47 Counties, Rural/Urban
Sampling Frame	NASSEP V (5,360 Clusters)
Sample Size & Allocation	
National	24,000 Households (2,400 Clusters)
Rural	14,120 Households (1,412 Clusters)
Urban	9,880 Households (988 Clusters)

9

Kenya, 2015

Kenya Integrated Household Budget Survey (KIHBS)

1.10 Survey response rates

The survey achieved high sample response rates. Nationally, 91 per cent of the sampled households participated and completed questionnaires. As shown in Table 1.2, from 23,852 households that were sampled for the survey, a total of 21,773 households were successfully interviewed. The response rate for rural households was higher (93.6%) compared to that of urban households (88.0%). Part of the non-response was due to non-coverage of 13 clusters spread across different counties occasioned by either insecurity or non-availability of households due to movement of populations in nomadic areas

Table 1.2: Response rates

	Residence		
Result	Urban	Rural	Total
Households selected	9,870	13,982	23,852
Households interviewed	8,681	13,092	21,773
Household response rate	88.0	93.6	91.3

C4D2 TRAINING 10

10

Documentation on fieldwork

what to include

- Training
  - Calendar
  - Quizzes
  - Evaluation forms and selection procedures
- Composition and territorial deployment of the field teams
- Dates of field work
- Problems encountered
- Changes to field procedures
- Supervision forms
- Non-response rates, by interviewer

C4D2 TRAINING 11

11

Uganda, 2016/17

National Household Survey

1.6.3 Fieldwork

A centralized approach to data collection was employed through which 13 mobile field teams grouped at the UBOS headquarters were deployed to the different sampled areas. Each team comprised one field supervisor, three or four enumerators and a driver. The field staff were recruited based on fluency of the local language spoken in the respective region of deployment while the supervisors were balanced between males and females. Prior to the deployment of fieldwork teams, ten listing teams each comprising of a team leader and two listers were constituted to update the number of households within the sampled EAs.

At the headquarters, a team of regional and senior supervisors undertook several other survey activities in line with the survey including data scrutiny, field monitoring, coordination and supervision among others. The field data collection commenced at the end of June 2016 and was completed in June 2017. Fieldwork was carried out in 12 separate trips, between which teams met at the headquarters for refresher training and debriefing sessions. During the meetings, the main issues discussed included logistical and data collection challenges which were resolved instantly.

C4D2 TRAINING 12

12

Definition of economic concepts and aggregates

an example from South Africa (p.125)

Non-durable goods – Household items that do not last long, for example food and personal care items. Households usually acquire these items on a daily, weekly or monthly basis.

Non-poor – Population living above a designated poverty line.

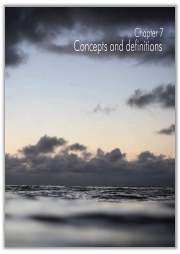
Own production – Own production is the activity of producing goods that the household can consume or sell in order to supplement the household income. Many households – especially low-income households – need to grow food items such as vegetables, mealies, etc., or to keep chickens or livestock to consume and/or sell so that they can provide more adequately for themselves.

Payment approach – An approach taking into account the total payment made for all goods and services in a given period, whether the household has started consuming them or not.

Poor – Population living below a designated poverty line.

Poverty gap – This provides the mean distance of the population from the poverty line (this is also referred to as P_1).

Poverty headcount – This is the share of the population whose income or consumption is below the poverty line; that is, the share of the population that cannot meet its basic needs (this is also referred to as P_0).



Poverty trends in South Africa
An examination of absolute poverty between 2006 and 2015

C4D2 TRAINING 13

13

2. Descriptive statistics

C4D2 TRAINING 14

14

How to describe data effectively

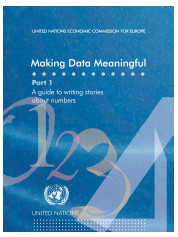
1) Text
2) Tables
3) Graphs

C4D2 TRAINING 15

15

Text

- Effective writing complements good tables and graphs
- Today we focus on the latter: writing deserves a separate discussion
- A useful reference



16

Tables

- Tables are omnipresent in data dissemination reports
- Often used when describing two variables jointly (two-way tables), e.g. income by region, population by age…

17

What do you think of this table?

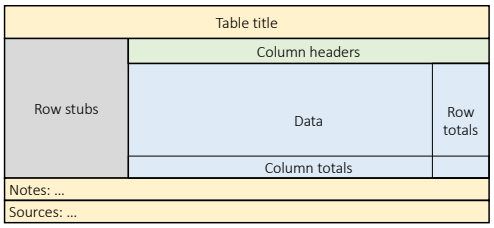
Income inequalities by area of residence and geopolitical Zone for 2001 and 2010

		2004	2010	% change from 2004 to 2010
National		0.4296	0.447	4.1
Area of Residence	Rural	0.4239	0.4334	2.2
	Urban	0.4154		4.2
Geo-political zones				
	1 South South	0.3849	0.434	12.8
	2 South East	0.376	0.4442	18.1
	3 South West	0.4088	0.4097	0.2
	4 North Central	0.4459	0.422	-5.4
	5 North East	0.4114	0.4468	8.6
	6 North West	0.4028	0.4056	0.7

18

Table elements

"The sandwich"



C4D2 TRAINING 19

19

What makes a good table

Golden rule #1

Express contents clearly

1. The table title should answer the questions "what", "where" and "when", but still be concise
2. Tables should be self-contained: use notes to clarify definitions, abbreviations, etc.
3. Percentage distributions of discrete variables should be clearly identified as either percentages of households or percentages of the population
4. Row and column totals should be reported, when they identify a marginal distribution

C4D2 TRAINING 20

20

What makes a good table

Golden rule #2

Reduce clutter

1. Avoid unnecessary colors, repetitions (*e.g.* use % or $ just once, in the title, rather than throughout the table)
2. Precision of numbers: do not present too many significant digits. Percentages: one decimal digit is usually enough. Numbers with four or more digits: no decimals at all. Large numbers: express them in thousands or millions
3. Be mindful of spacing and alignment

C4D2 TRAINING 21

21

What's wrong with this table?

UNECE (2009: 12)

Final energy consumption by sector - Percentages

	1980	1985	1990	1995	2000	2002	2003
Transport	27.81	27.92	28.24	31.12	36.82	39.48	39.13
Residential	31.11	33.91	30.41	27.61	24.33	23.71	23.97
Industry	31.47	27.21	23.86	22.11	21.41	19.53	18.78
Agriculture	n/a	n/a	3.51	3.7	3.11	2.91	2.82
Services	9.61	10.96	13.98	15.46	14.33	14.37	15.3

22

A possible table redesign

UNECE (2009: 12)

Share of total energy consumption, by sector (in percent)
Ireland, 1980-2003

	1980	1985	1990	1995	2000	2002	2003
Transport	27.8	27.9	28.2	31.1	36.8	39.5	39.1
Residential	31.1	33.9	30.4	27.6	24.3	23.7	24.0
Industry	31.5	27.2	23.9	22.1	21.4	19.5	18.8
Agriculture	n/a[1]	n/a[1]	3.5	3.7	3.1	2.9	2.8
Services	9.6	11.0	14.0	15.5	14.4	14.4	15.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

[1] Data on energy consumption for the agricultural sector was not collected until 1990.
Source: Department of Public Enterprise, Ireland

Title is clear, descriptive and self-contained

Formatting is not cluttered
Row stubs and column headers are emphasized
Useless decimals are omitted

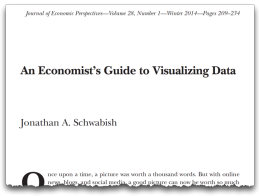
Column totals clarify content of the table

Note clarifies meaning of "n/a"; data source is clearly stated

23

Graphs

- In many cases, presentation of data can be made more interesting and intuitive by using graphs or charts rather than tables.
- Many of the "golden rules" that help make better tables also apply to graphs.



Journal of Economic Perspectives—Volume 28, Number 1—Winter 2014—Pages 209–234

An Economist's Guide to Visualizing Data

Jonathan A. Schwabish

Once upon a time, a picture was worth a thousand words. But with online news, blogs, and social media, a good picture can now be worth so much

24

What makes a good graph

Golden rule #3

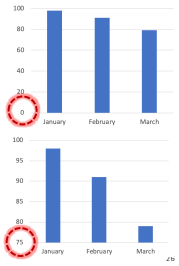
Express contents clearly

1. A good graph title answers the same questions as a good table title
2. Graphs should be self-contained too (use notes)
3. Explain encoding: always label axes and data series clearly
4. Avoid visualizations that mislead the eye: two notorious "sins" are bar charts with a nonzero baseline, and 3D pie charts

25

Bar charts with nonzero baseline

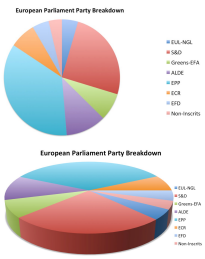
- Bar charts rely on bar length to show data: compare lengths to compare values
- Shifting the baseline distorts the visual: a value twice as high no longer corresponds to a bar twice as long
- Graphs on the right show the same data, but appear very different



26

3D pie charts

- Pie charts encode data in the area of each slice: larger slice equals higher share
- A 3D pie chart distorts angles, making the slice that is "closer" to the viewer appear larger than it actually is
- This visualization can mislead viewers, and should be avoided



Source: Hickey, Business Insider, Jun. 17, 2013

27

What makes a good graph

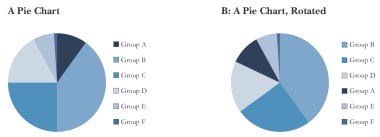
Golden rule #4

Reduce clutter

1. Again, avoid unnecessary colors and decorative elements that obfuscate the message of the graph
2. Precision of numbers: same recommendations as for tables
3. Do not crowd graph with too many data points: viewer should be able to understand the message of the graph easily, without having to parse too much visual information (if that is the issue, select a subset of relevant values, or consider using a table instead)

28

On pie charts

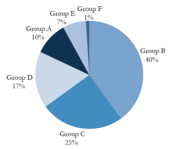


"Because pie charts force readers to make comparisons using the areas of the slices or the angles formed by the slices—something that our visual perception does not accurately support — they are not an effective way to communicate information" Schwabish (2014: 223)

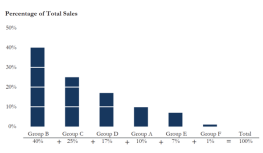
29

Graph redesign

A pie chart, labeled



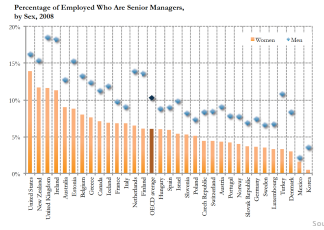
Pie chart alternative:
a bar or column chart



Source: Schwabish (2014: 223)

30

What's wrong with this graph?



Source: Schwabish (2014: 218)

C4D2 TRAINING

31

31

Graph redesign

Source: Schwabish (2014: 220)

C4D2 TRAINING

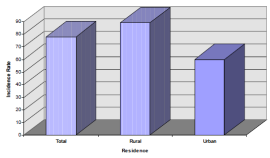
32

32

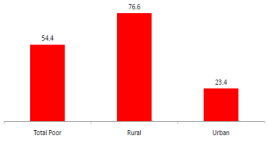
What's wrong with this graph?

Zambia, Living conditions monitoring survey report 1996 and 2015

Incidence of poverty rural/urban, 1996



Incidence of poverty rural/urban, 2015

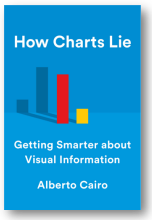


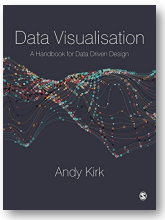
C4D2 TRAINING

33

33

Useful references





34

3. Inequality and poverty

35

Overview

- Tips for presentation of generic summary statistics still apply
- There are a few additional points to be made specifically about presenting results on poverty and inequality:

a. Popular measures and graphics (from lectures 13 and 14)

b. Best practices for making comparisons

36

Suggested inequality measures

Malawi poverty assessment IHS2 IHS3, Inequality indices

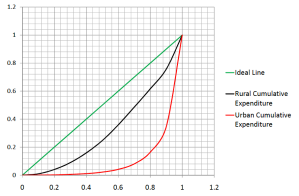
- Gini is so prevalent internationally that it cannot be omitted, the rest is extra credit

	Malawi		Urban		Rural	
	2004	2010	2004	2010	2004	2010
GE (-1)	0.28	0.41	0.48	0.54	0.21	0.28
MLD (GE(0))	0.25	0.34	0.39	0.41	0.19	0.23
Theil index (GE(1))	0.31	0.42	0.44	0.47	0.20	0.25
GE (2)	0.58	0.96	0.73	0.88	0.29	0.38
Gini	0.39	0.45	0.48	0.49	0.34	0.38

37

Suggested inequality charts: The Lorenz Curve

Zambia, Living conditions monitoring survey report 2015



- Much like the Gini index, the Lorenz curve is a staple when reporting about inequality
- Especially useful when comparing inequality across sub-populations (Lorenz dominance)

38

Suggested poverty measures

2015/16 Kenya Integrated Household Budget Survey (KIHBS)

- FGT, the rest is extra credit

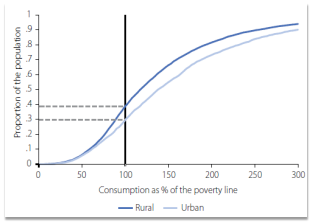
Table 4.3: Overall Poverty Estimates (Individual) by Residence and County, 2015/16

Residence / County	Headcount Rate (%)	Distribution of the Poor (%)	Poverty Gap (%)	Severity of Poverty (%)	Population ('000)	Number of Poor ('000)
National	36.1	100.0	10.4	4.5	45,371	16,401
Rural	40.1	71.3	11.5	5.0	29,127	11,687
Peri-Urban	27.5	5.6	6.9	2.6	3,340	920
Core-Urban	29.4	23.1	8.9	3.9	12,905	3,795

39

Suggested poverty charts: overlayed CDFs

Kenya gender and poverty assessment 2015/16



- Stochastic dominance analysis is useful when comparing poverty in sub-populations
- Here one CDF lies below the other, at any level of consumption
- Urban First Order Stochastically Dominates (FODs) rural
- Interpretation: the incidence of absolute poverty is lower in urban areas, irrespective of the level chosen for the poverty line.

40

Making comparisons

- Many audiences (policy makers, general public) are especially interested in comparisons of poverty and inequality, over time or across regions
- Poverty and inequality trends are among the most visible and impactful results to emerge during dissemination
- Comparability of underlying data and methods is key: if processes that led up to estimates differ, comparison is invalid
- Being transparent on comparability is key!

41

Changes in data and methodology

- Comparability of data and methods underlying the estimates that are being presented is key
- If processes that led up to estimates differ, comparison is invalid and readers may be misled
- Minimize incomparability
- If some discrepancies remain, be fully transparent

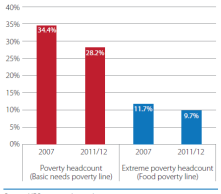
42

Tanzania, 2012

Poverty Assessment

The analysis of the poverty trend is challenged by changes in the HBS design, but the adjustments made to counter the change in design support the decline of poverty. Assessing the changes in poverty levels over time is subject to issues of comparability stemming from changes in the survey design and methodological improvements implemented during the 2011/12 HBS. These issues were addressed using different methods, including the reevaluation of the consumption aggregates for HBS 2007 using the same approach as in 2011/12, as well as nonparametric and parametric imputation procedures. The different

Figure ES.1 Poverty and Extreme Poverty Incidence



Source: HBS 2007 and 2011/12.

C4D2 TRAINING

43

43

Tanzania, 2012

Poverty Assessment, HBS 2007 and 2011/12 recall modules

	HBS 2011/12 Recall period (months)			HBS 2007 Recall period (months)		
Consumption and expenditure categories	1	3	12	1	3	12
Clothing and footwear (COICOP 3)			X			X
Housing and utilities (COICOP 04 + selected other)						
Rents	X			X		
Utilities	X					X
Energy		X				X
Building maintenance			X			X
Housing equipment (COICOP 05)						
Household durables, furniture and furnishings			X			X
Small household appliances		X				X
Expenditures on domestic workers	X					X
Health expenditures (COICOP 06)	X					X

C4D2 TRAINING

44

44

The importance of uncertainty

- Poverty calculations are based on a sample of households, and samples carry a margin of error in representing the population
- Standard errors should always be estimated along with poverty point estimates
- Crucial when making comparisons (over time, across regions): poverty changes should not be taken at the face value
- Note: probability weighting, clustering, and stratification, are survey design features which must be taken into account when estimating standard errors.

C4D2 TRAINING

45

45

Ethiopia, 2015

Household Income Consumption & Expenditure

Table 9: Poverty indices in 2015/16

	Estimate	Std. Err.	[95% Conf.	Interval]
Poverty head count index	0.235	0.008	0.220	0.250
Poverty gap index	0.067	0.003	0.061	0.073
Poverty severity index	0.028	0.002	0.024	0.031
Food poverty head count index	0.248	0.008	0.233	0.263
Food poverty gap index	0.067	0.003	0.061	0.073
Food poverty severity index	0.027	0.002	0.024	0.030

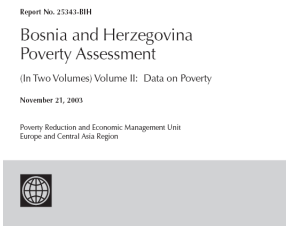
Source: computed from the 2015/16 HICE survey data

C4D2 TRAINING 46

46

Sensitivity Analysis

Bosnia and Herzegovina, 2003 (vol. II)



Report No. 25343-BIH

Bosnia and Herzegovina
Poverty Assessment

(In Two Volumes) Volume II: Data on Poverty

November 23, 2003

Poverty Reduction and Economic Management Unit
Europe and Central Asia Region

C4D2 TRAINING 47

47

Sensitivity Analysis

Excerpt from the table of contents

C4D2 TRAINING 48

48

Sensitivity Analysis

Conclusions

Table 6.2 Key characteristics of poverty and its robustness to measurement assumptions.

Characteristics of poverty	Baseline, consumption per capita	OECD I scale	OECD II scale	Higher poverty line	Lower poverty line	Expenditure per capita
Mixed (semi-urban) municipalities in RS	yes	yes	yes	yes	yes	yes
Rural municipalities in FBiH	yes	yes	no	no	yes	no
IDPs and Refugees	yes	yes	yes	yes	yes	yes
Households headed by persons with low education (primary or less)	yes	yes	yes	yes	yes	yes
Households headed by persons with education above secondary	no	no	no	no	no	no
Unemployed (ILO) and inactive adults	yes	yes	yes	yes	yes	yes
Employed according to registration	no	no	no	no	no	no
Registered unemployed	yes	yes	yes	yes	yes	yes
Household headed by elderly	no	no	no	no	no	no
Larger households	yes	yes	no	yes	yes	yes

Source: Staff estimates based on BiH-LSMS 2001.

C4D2 TRAINING 49

49

Lessons learned

1. Explain clearly what your table or graph is showing (titles, labels, notes…)
2. Only if point 1 is checked, reduce clutter: keep frills to a minimum
3. When showing results on inequality and poverty, include the shortlist of key measures and graphs indicated in this lecture, which the international community has come to expect
4. Comparisons: document changes in data and methodology, and include measures of uncertainty of estimates whenever possible

C4D2 TRAINING 50

50

References

Required readings

Glewwe, P., and Levin, M. (2005). Presenting simple descriptive statistics from household survey data. In UN, Household Sample Surveys in Developing and Transition Countries. Studies in Methods Series F No. 96

Schwabish, J. A. (2014). An economist's guide to visualizing data. Journal of Economic Perspectives, 28(1), 209-34.

Suggested readings

Dupriez, O., & Boyko, E. (2010). Dissemination of Microdata Files: Principles Procedures and Practices. International Household Survey Network.

UNECE (2009). Making Data Meaningful, Part 1: A Guide to writing stories about numbers. United Nations, Geneva.

UNECE (2009). Making Data Meaningful, Part 2: A Guide to presenting statistics. United Nations, Geneva.

Suggested websites and blogs
(tutorials, tips, data visualization examples)

Alberto Cairo: thefunctionalart.com

Stephanie Evergreen: stephanieevergreen.com

Nathan Yau: flowingdata.com

Jonathan Schwabish: policyviz.com

C4D2 TRAINING 51

51

Thank you for your attention

C4D2 Training 52

52

Homework

C4D2 Training 53

53

Exercise 1 – Engaging with the literature

- The dissemination of microdata often (but not always) accompanies the dissemination of findings and summary statistics from a survey
- Summarize the discussion of the pros and cons of data dissemination in Dupriez et al. (2010) p. 16-23
  http://ihsn.org/sites/default/files/resources/IHSN-WP005.pdf

C4D2 Training 54

54

Exercise 2 - Standard Errors

Poverty analysis of the integrated household survey in The Gambia 2003 (p49-50)

- Compare point estimates with interval estimates, assuming a 95% confidence level, and briefly comment on results

Table 5: Poverty by area

Poverty	Area	Estimate	Std. Err.
Head count index	Urban	33.4	3.0
	Rural	60.6	2.8

Table 7: Poverty by strata

Poverty	Strata	Estimate	Std. Err.
Head count index	*Banjul Urban*	6.6	4.0
	KMC urban	32.1	4.0
	Brikama Urban	41.9	9.4
	Brikama Rural	50.1	4.8
	Mansakonko Urban	65.8	12.8
	Mansakonko Rural	55.9	8.3
	Kerewan Urban	42.7	10.3
	Kerewan Rural	67.1	5.9
	Kuntaur Urban	57.1	9.2
	Kuntaur Rural	91.9	3.8
	Janjangbureh Urban	53.0	22.4
	Janjangbureh Rural	63.0	8.4
	Basse Urban	44.4	7.9
	Basse Rural	63.3	8.0

C4D2 TRAINING

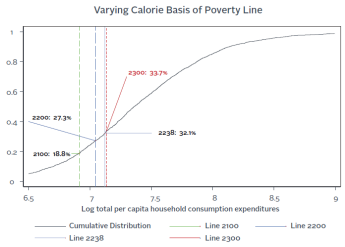
55

55

Exercise 3 - Sensitivity analysis

Myanmar Poverty and Living Conditions Survey 2015

- Briefly comment on the robustness of the poverty line to different calorie norms.



C4D2 TRAINING

56

56

Acknowledgments

The course was prepared by Giovanni Vecchi (U. of Rome Tor Vergata) and Giulia Mancini (U. of Rome Tor Vergata), with a core team comprising Nicola Amendola and Sédi Anne-Boukaka (U. of Rome Tor Vergata).

The work was conducted under the supervision of Gero Carletto, Michelle Jouvenal, Shelton Kanyanda, and Alberto Zezza (World Bank), who provided comments throughout the drafting process.

The team relied on advice and comments from many colleagues, who also contributed their own teaching material – Giovanni D'Alessio, Romina Gambacorta, Giuseppe Ilardi, Valentina Michelangeli, Andrea Neri (Bank of Italy), Gero Carletto, Dean Joliffe, Shelton Kanyanda, Talip Kilic, Heather Moylan, Diane Steele, Alberto Zezza (World Bank), Federico Polidoro, Paolo Consolini, Valeria De Martino, Gabriella Donatiello (Istat), and Piero Conforti (FAO).

C4D2 TRAINING

57

57