Tulane University

Appraising World Income Inequality Databases: An Overview

Nora Lustig
Tulane University, CGD, IAD

UNU-WIDER Conference
"Inequality – Measurement, Trends, Impacts and Policies"
Helsinki, Finland, September 5, 2014

Special Issue of JOEI

"Appraising World Income Inequality Databases"

Francisco Ferreira (World Bank) and Nora Lustig (Tulane University), editors

Research Assistant: Dan Teles (Tulane University)

- Reviewed Databases: 9
- To be published (online) end of 2014/early 2015
- Today's presentations: A Preview (preliminary)

Assessing Inequality Databases

- Why an assessment is not only desirable but necessary?
- Which databases were included?
  - Microdata: 6
  - Secondary Source: 2
  - Imputations: 1

The Six Microdata-based Databases and their Reviewers

1. CEPALStat (UNECLAC): François Bourguignon (Paris School of Economics)

2. IDD (Income Distribution Database; OECD): Leonardo Gasparini and Leopoldo Tornarolli (University of La Plata)

3. LIS: Martin Ravallion (Georgetown University)

4. POVCAL/WDI (World Bank): Tim Smeeding and Jonathan Latner (University of Wisconsin)

The Six Microdata-based Databases and their Reviewers

5. SEDLAC (Socioeconomic Database for LAC; CEDLAS, Univ de La Plata and the World Bank): François Bourguignon (Paris School of Economics)

6. WTID (World Top Incomes Database; Atkinson, Piketty, Saez and Alvaredo): Andrea Brandolini (Bank of Italy

From 1 – 5: microdata Household Surveys

WTID: microdata mainly tax returns (complete, samples or tabulations)

The Two Secondary Source-based and the One Imputations-based Databases and their Reviewers

Secondary Source-based:

1. **ATG (All The Ginis; Branko Milanovic):** Tim Smeeding and Jonathan Latner (University of Wisconsin)
2. **WIID (World Income Inequality Database)**: Stephen Jenkins (London School of Economics)

Imputations-based:

1. **SWIDD (Standardized World Income Inequality Database)**: Stephen Jenkins (London School of Economics)

Where are these databases produced institutionally?

- All the Ginis is produced privately by Branko Milanovic (presently at LIS)
- CEPALSTAT is produced by the United Nations Economic Commission for Latin America and the Caribbean; based in Santiago, Chile
- IDD is produced by OECD; Paris, France
- LIS Key Figures are produced by LIS, a private organization whose current director is Janet Gornick; based in Luxembourg and New York

Where are these databases produced institutionally?

- POVCAL is produced by World Bank; Washington, DC
- SEDLAC is a "joint venture" of CEDLAS (an Argentine research center at Univ. de La Plata) and the World Bank
- SWIID is produced by Frederick Solt, Assistant Professor, Dept. of Political Science, University of Iowa
- WIID is produced by UNU-WIDER; Helsinki, Finland
- WTID is produced by Facundo Alvaredo, Antony Atkinson, Thomas Piketty and Emmanuel Saez; housed at the Paris School of Economics

Databases not included and worth mentioning

- University of Texas Income Project (UTIP)
- The Gini Project
- Commitment to Equity (CEQ); based at Tulane University
- Global Consumption and Income Project (GCIP)

Gini Coefficient Frequencies in Primary Source Datasets (CEPAL, LIS, SEDLAC, OECD IDD, and WDI/POVCAL)

Region	Number of Country-Years with Primary Source Data	Total Number Primary Source Datapoints	Earliest Observation	Most Recent Observation
East Asia and Pacific	120	123	1981	2011
Eastern Europe and Central Asia	301	334	1984	2011
Latin America and Caribbean	378	832	1974	2013
Middle East and North Africa	52	54	1979	2010
South Asia	39	39	1978	2012
Sub-Saharan Africa	140	140	1980	2011
Western Europe and North America	316	403	1967	2010
Grand Total	**1346**	**1925**	**1967**	**2013**

NOTE: Statistics as of January 2014

Things we want to know about a database

Most users just want to know:

- Inequality indicators
- Country coverage
- Period coverage

Things we want to know about a database

More sophisticated users also want to know:

- Welfare indicator:
  - Per capita or equivalized
  - Income- or consumption-based
  - Total or monetary
  - Before or after taxes and/or transfers
- Statistical significance

Things we want to know about a database

- Are income concepts homogenized for comparability
- Were indicators calculated from unit records or grouped data
- Are regional price differentials taken into account
- What is the definition of household (e.g., domestic servants and boarders)

Things we want to know about a database

- Data adjustments; can they be replicated:
  - correction for under-reporting
  - top coding
  - treatment of extreme values and zeros or negative incomes
- Information on the survey (sample design, questions, recall periods, etc.) and their comparability across countries and over time
- Is it possible to have access to the microdata

Description of the Datasets	Group 1: Datasets that Calculate Indices with Microdata					
Dataset	CEPALStat	LIS	IDD	SEDLAC	WDI	WTID
Inequality Indicators (Gini (G), Theil (T), Atkinson (A), Others (O))	G,T,A, O	G,T,A, O	G,O	G,T,A, O	G,O	O
Statistical Significance Indicators (i.e., standard errors or confidence intervals) (Always (A), Sometimes (S), Never (N))	N	N	S	A	A	N
Is data comprised of individual observations (I) or grouped data (G)?	I	I	G	I	Both	Both

Dataset	CEPAL	LIS	OECD	SEDLAC	WDI	WTID
Description of Welfare Concept						
Income (I) or consumption (C)	I	I	I	I	varies	I
Monetary (M) or total (T)? If 'total', does it include autoconsumption (Yes(Y)/No(N)) , imputed rent (Yes(Y)/No(N))?	T(Y,Y)	T(Y,N)	M	T(Y,Y)	varies	varies
Includes estimates before taxes and transfers? (Yes(Y)/No(N))	NS	N	Y	N	NS	Y
Includes estimates after taxes and transfers? (Yes(Y)/No(N))	NS	Y	Y	Y*	NS	N
Unit of analysis: per individual (I), per household (H), per equivalence scales (E)?	I	E	E	I & E	I	varies
Are differences in prices by region (e.g., rural urban, etc.) accounted for?(Yes(Y)/No(N))	???	N	N	Y	varies	NS

*It is assumed that individuals report income after taxes for the employed and before taxes for the self-employed

Dataset	CEPAL	LIS	OECD	SEDLAC	WDI	WTID
Adjustments to the original data source (e.g. for harmonization purposes)						
Correction for under-reporting (Yes(Y)/No(N))	Y	N	varies	N	N	varies
Is documentation sufficient to replicate results? (Yes(Y)/No(N))	N	N/A	N	N/A	N/A	Y
Adjustment for top coding? (Yes(Y)/No(N))	N	Y	N	N	NS	N/A
Elimination of extreme values (Yes(Y)/No(N))	N	N	varies	N	NS	N
Is access to microdata made available through the dataset provider? (Yes(Y)/No(N))	N	Y	N	N	N	N

Description of the Datasets	Group 2: Datasets that use Secondary Sources		Group 3: Imputed
Dataset	**ATG**	**WIID**	**SWIID**
Dataset Summary			
Inequality Indicators (Gini (G), Theil (T), Atkinson (A), Others (O))	G	G, O	G
Statistical Significance Indicators (i.e., standard errors or confidence intervals) (Always (A), Sometimes (S), Never (N))	N	N	A

Dataset	ATG	WIID
Adjusts primary source data?(Yes(Y)/No(N))	N	N
Is original source of data clearly noted?(Yes(Y)/No(N))	Y	Y
Are welfare concepts clearly noted? (Yes(Y)/No(N))	N	Y
If mulitiple datapoints are available for the same country and year, are some sources of data given priority?(Yes(Y)/ No(N))	Y	Y
If mulitiple datapoints are available for the same country and year,is a "first-best" datapoint selected? (Yes(Y)/No(N))	Y	N
Are databases that use secondary data sources in turn used as inputs? (Yes(Y)/No(N))	Y	N

Dataset: SWIID	
Methodology	
Is description of imputation methods sufficient to replicate?	Y
Has method been subject to scrutiny by experts in the field of imputation?	Not clear
Is there a systematic validation process in place with experts on countries/regions?	Not clear
Is it clear how the Gini coefficient for income before taxes and transfers is calculated?	N

Sources Used by Secondary Source Datasets: All the Ginis, SWIID, and WIID

	Secondary and Imputed Datasets		
	All the Ginis	**SWIID**	**WIID**
Sources Used			
Group 1: Datasets that Calculate Indices with Microdata			
CEPALSTAT		X	
Luxembourg Income Study (LIS)	X	X	X
OECD IDD			
Socio-Economic Database for Latin America and The Caribbean (SEDLAC)	X	X	X
World Development Indicators(WDI)	X	X	X
Group 2: Datasets that use Secondary Sources			
All the Ginis		X	
The Standardized World Income Inequality Database (SWIID)			
World Income Inequality Database (WIID)	X	X	

CEPALSTAT vs. SEDLAC

- Large overlap: 173 country-year combinations that appear in both datasets (out of 299 and 213)
- Both calculate Gini Coefficients directly from household survey microdata.

- **Important difference:** CEPALSTAT's corrects for underreporting

CEPALSTAT vs. SEDLAC

How likely is this difference of affecting our analysis of levels and trends in inequality in Latin America?

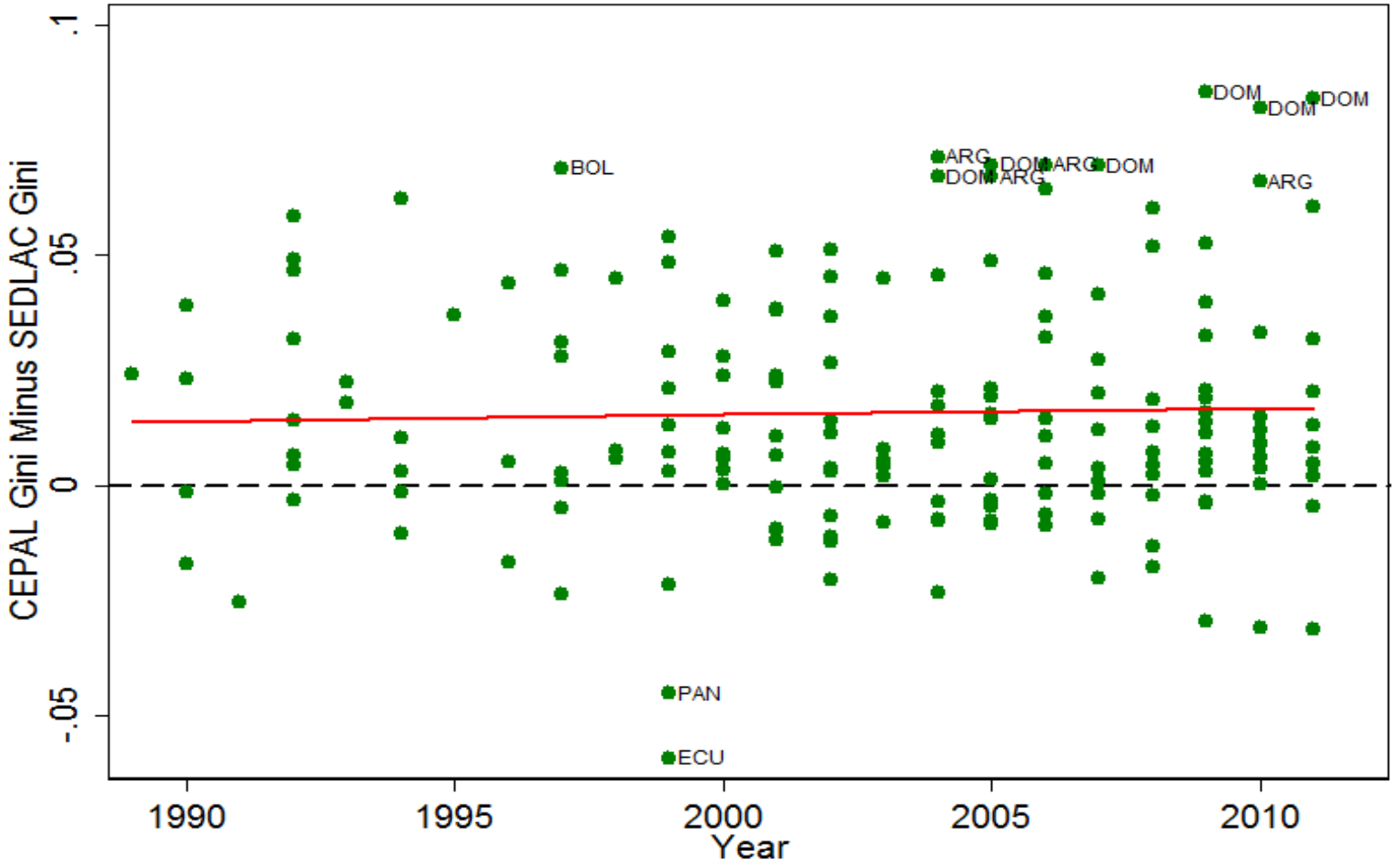
- Similar results in trends

- Data points are quite correlated (0.86)

CEPALSTAT vs. SEDLAC

- However, inequality levels—as expected—tend to be systematically and significantly higher in CEPALSTAT than in SEDLAC's, which does not correct for underreporting
- One additional problem of CEPALSTAT is that the correction method used to eliminate underreporting is not well documented and, therefore, cannot be replicated or compared with other approaches

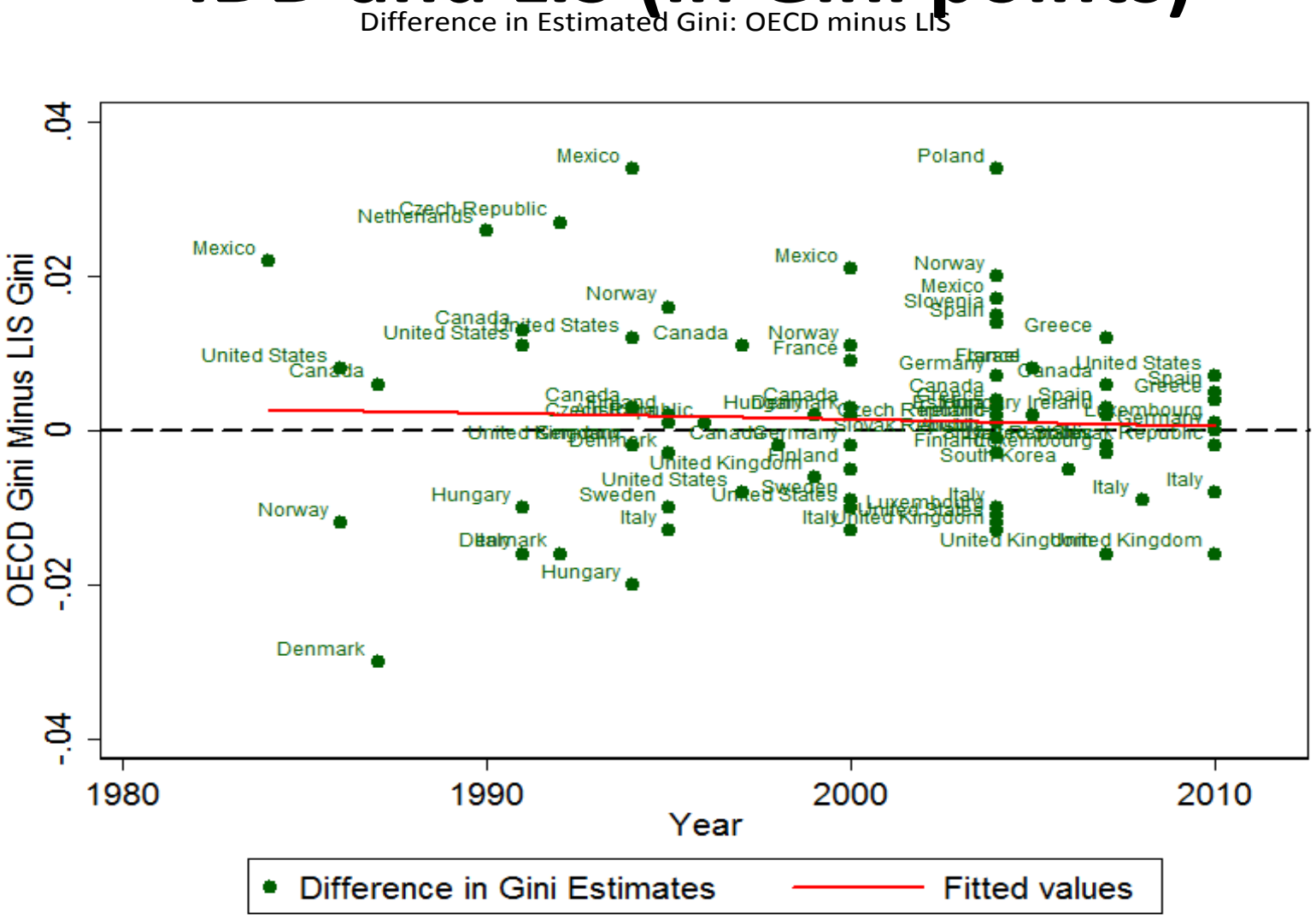
Difference in estimated Gini between CEPAL and SEDLAC (in Gini points)



LIS Key Figures vs. OECD IDD

- Large overlap: 79 country-year combinations that appear in both datasets (out of 203 and 326)
- Similar results in levels and trends.
  - Both datasets show a jump in inequality in Italy in the early 1990s, and fairly steadily increasing inequality in Germany, Israel, and the United States.
- Highly correlated (0.98)
- Nevertheless, when zooming in to a particular country/year, there can be important differences

Difference in estimated Gini between IDD and LIS (in Gini points)

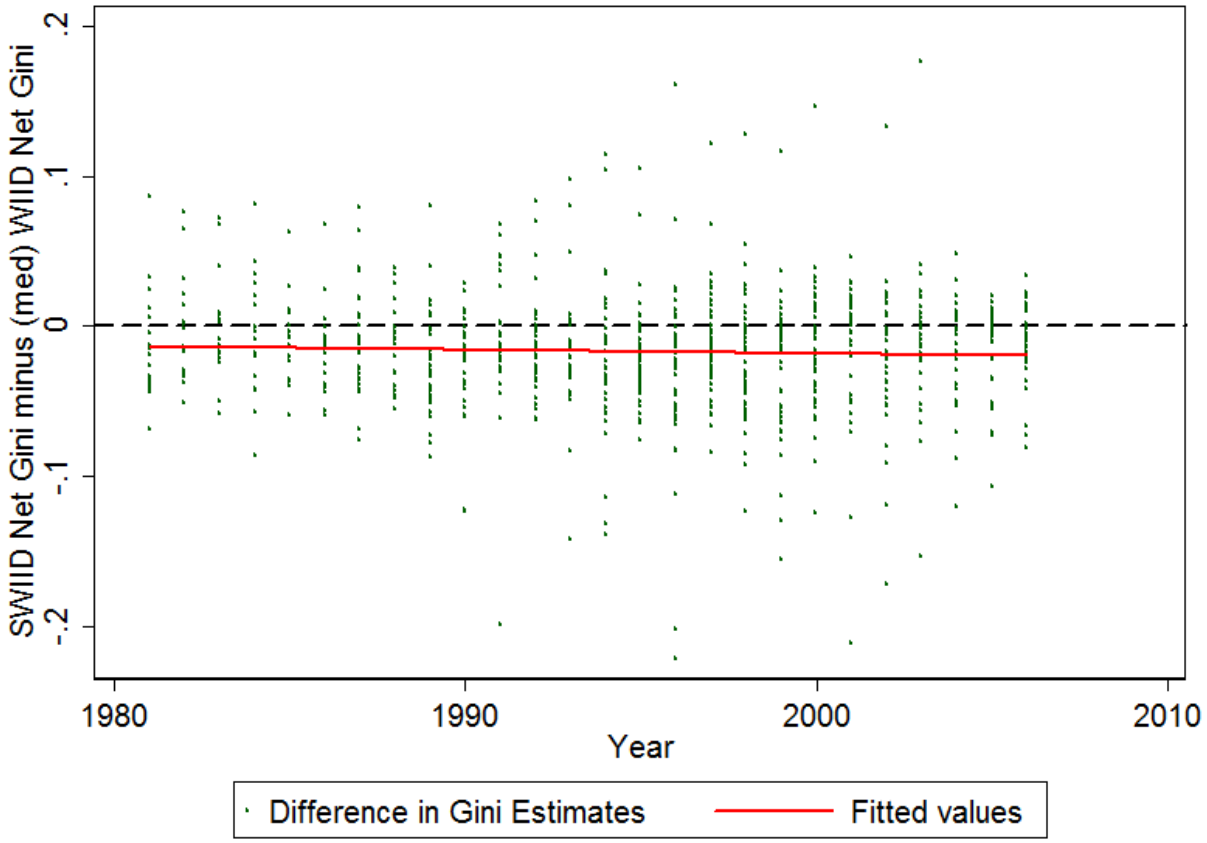


SWIID vs. Others

- What is the risk of making erroneous statements/inferences if we use the fully imputed dataset SWIID?
- While general trends tend to look fairly similar, there are important –devastating– differences when zooming in to a particular country-period

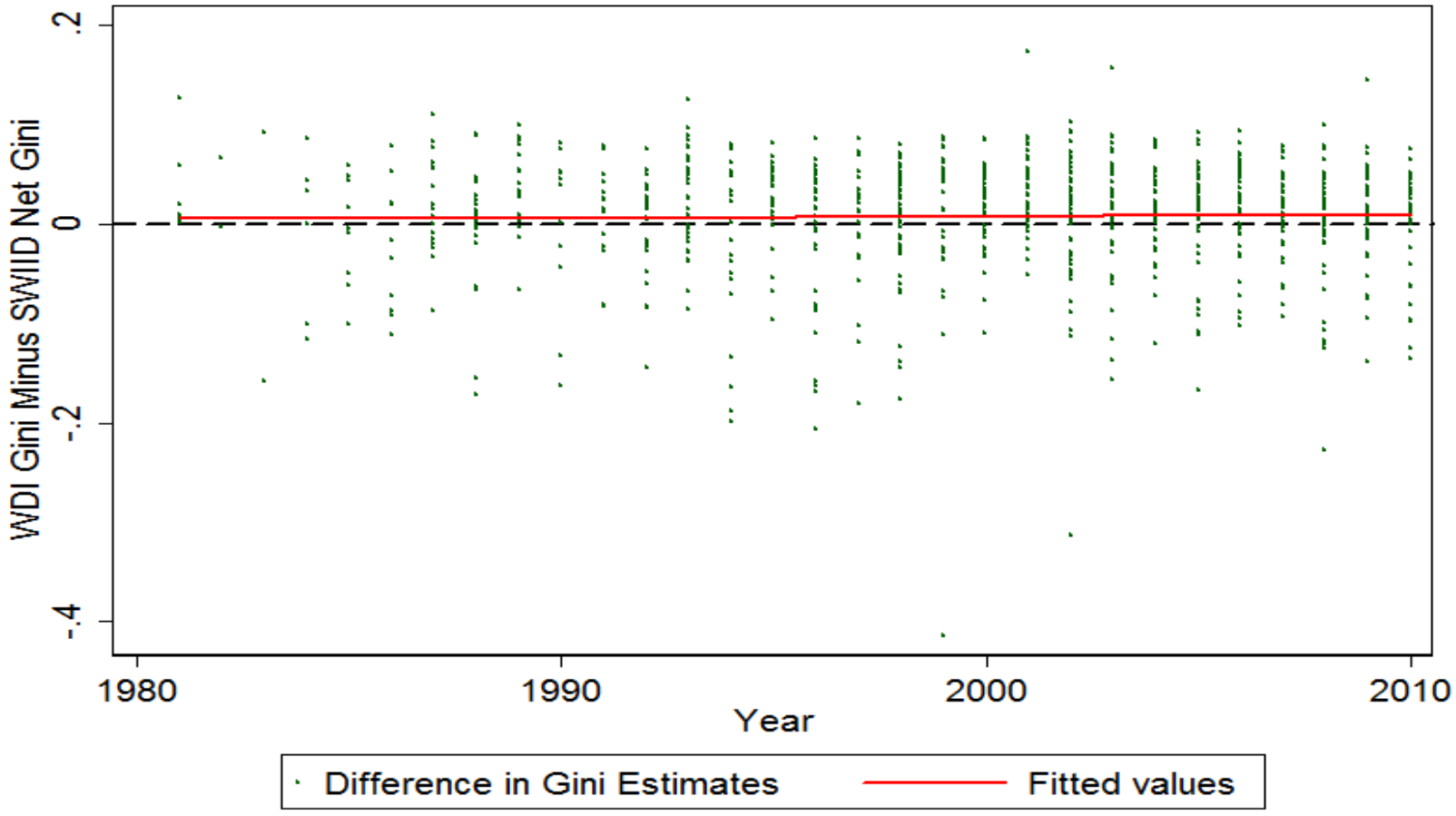
Difference in estimated Gini between SWIID and WIID (in Gini points)

Difference in Estimated Gini: SWIID Net Market Gini minus (median) WIID Net Gini



Difference in estimated Gini between POVCAL and SWIID (in Gini points)

Difference in Estimated Gini: POVCAL minus SWIID (for what SWIID calls Net Market Income)



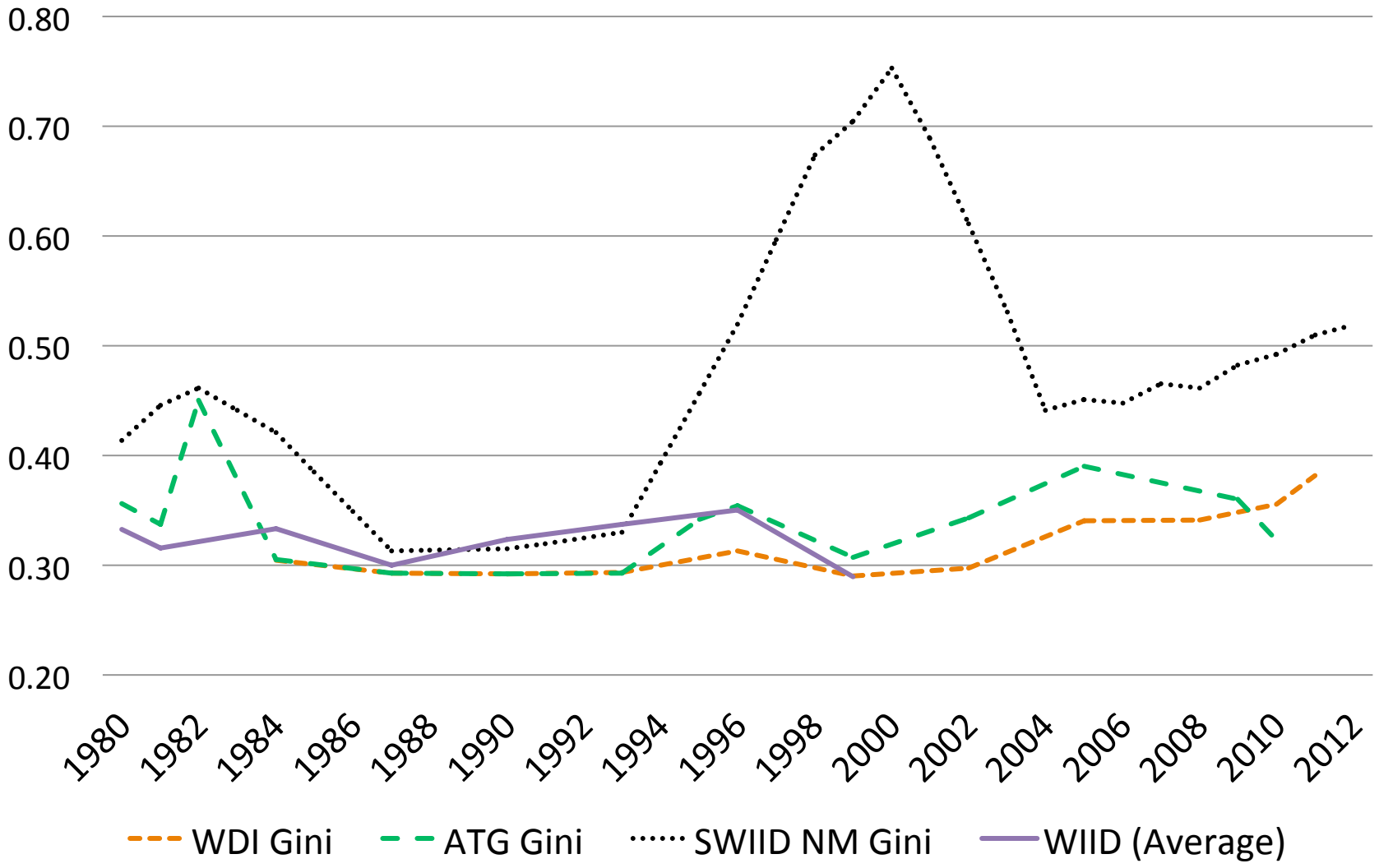
Has Inequality Declined in SSA?

The comparison that motivated the special issue:

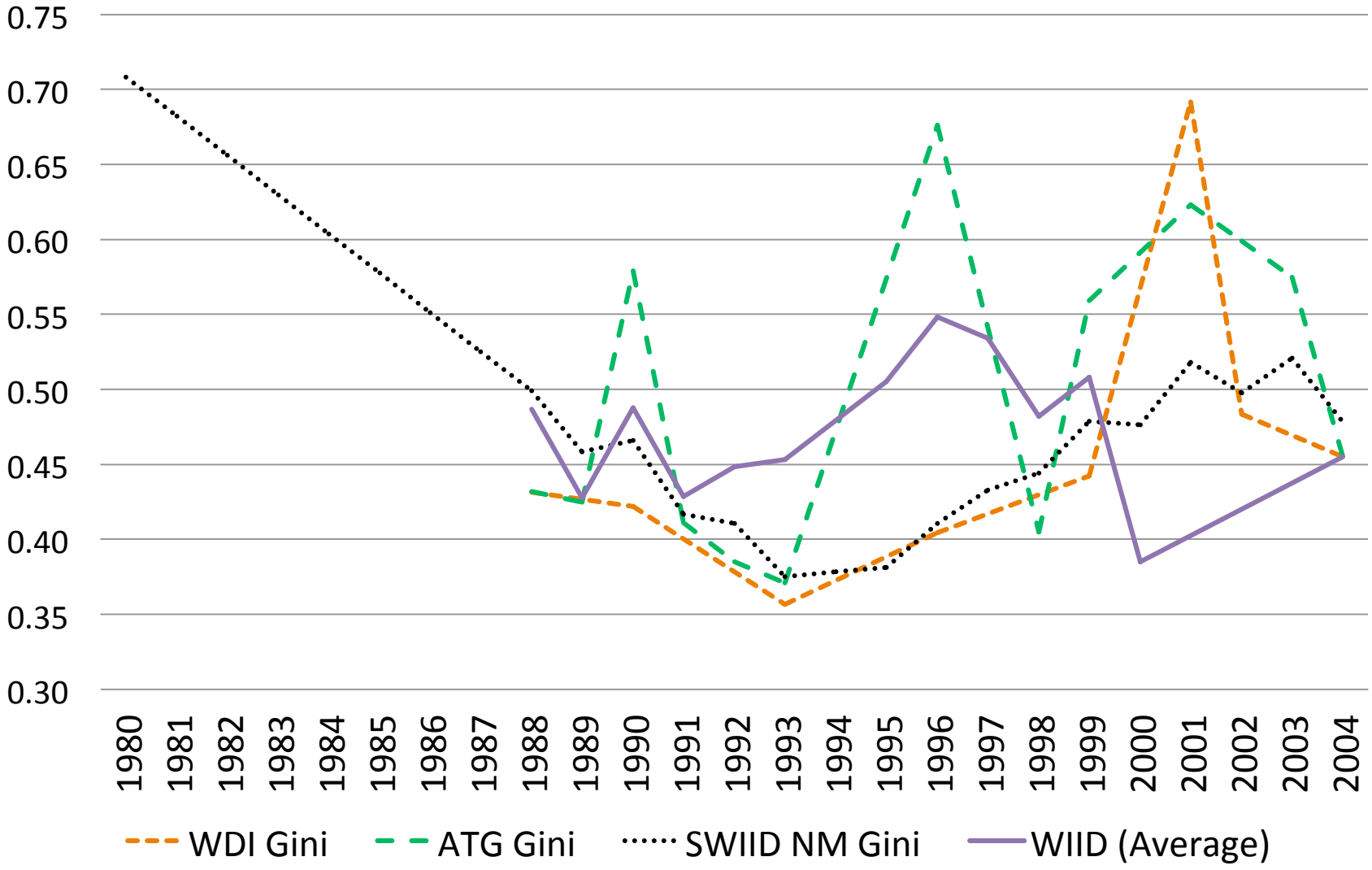
- POVCAL vs. the IMF Fiscal Monitor (Gini coefficients from SWIID): early 1990s with late 2000's
- In FOUR out of NINE cases in which IMF "found" that inequality declined, POVCAL's Ginis from microdata showed an increase

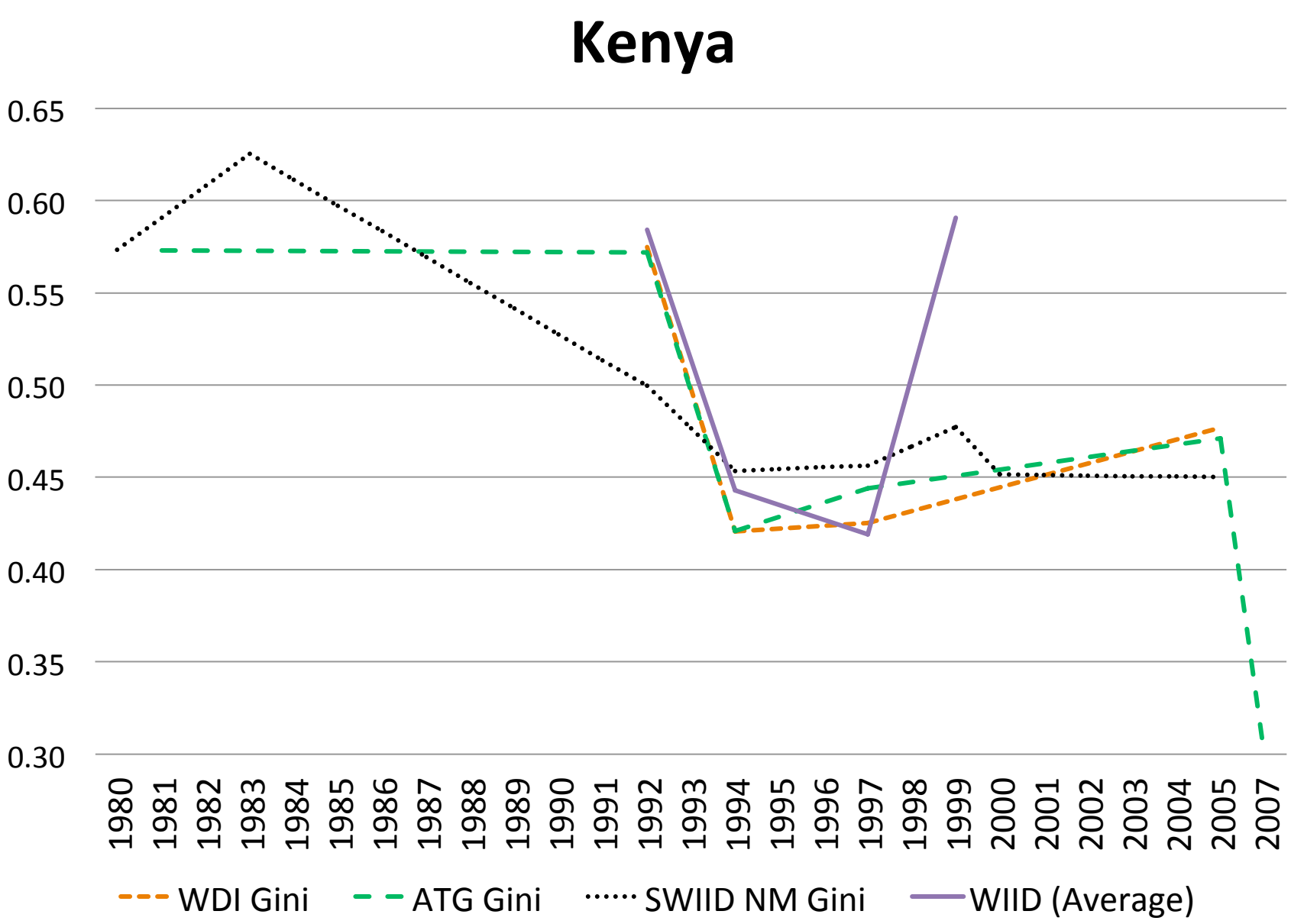
Country	Year	POVCAL	IMF Fiscal Monitor based on SWIID
Côte d'Ivoire	2008	41.5	45.3
Côte d'Ivoire	1993	36.9	40.3
Ghana	2005.5	42.8	40.1
Ghana	1991.5	38.1	37.7
Kenya	2005.4	47.7	46.1
Kenya	1994	42.1	52.3
Madagascar	2010	44.1	44.2
Madagascar	1993	46.1	45.2
Niger	2007.5	34.6	43.3
Niger	1992	36.1	44.8
Nigeria	2009.8	48.8	44.7
Nigeria	1992.3	45.0	49.5
Senegal	2005	39.2	37.2
Senegal	1991	54.1	45.1
Tanzania	2007	37.6	34.5
Tanzania	1991.9	33.8	37.6
Zambia	2006	54.6	49.5
Zambia	1993	52.6	63.1

Indonesia

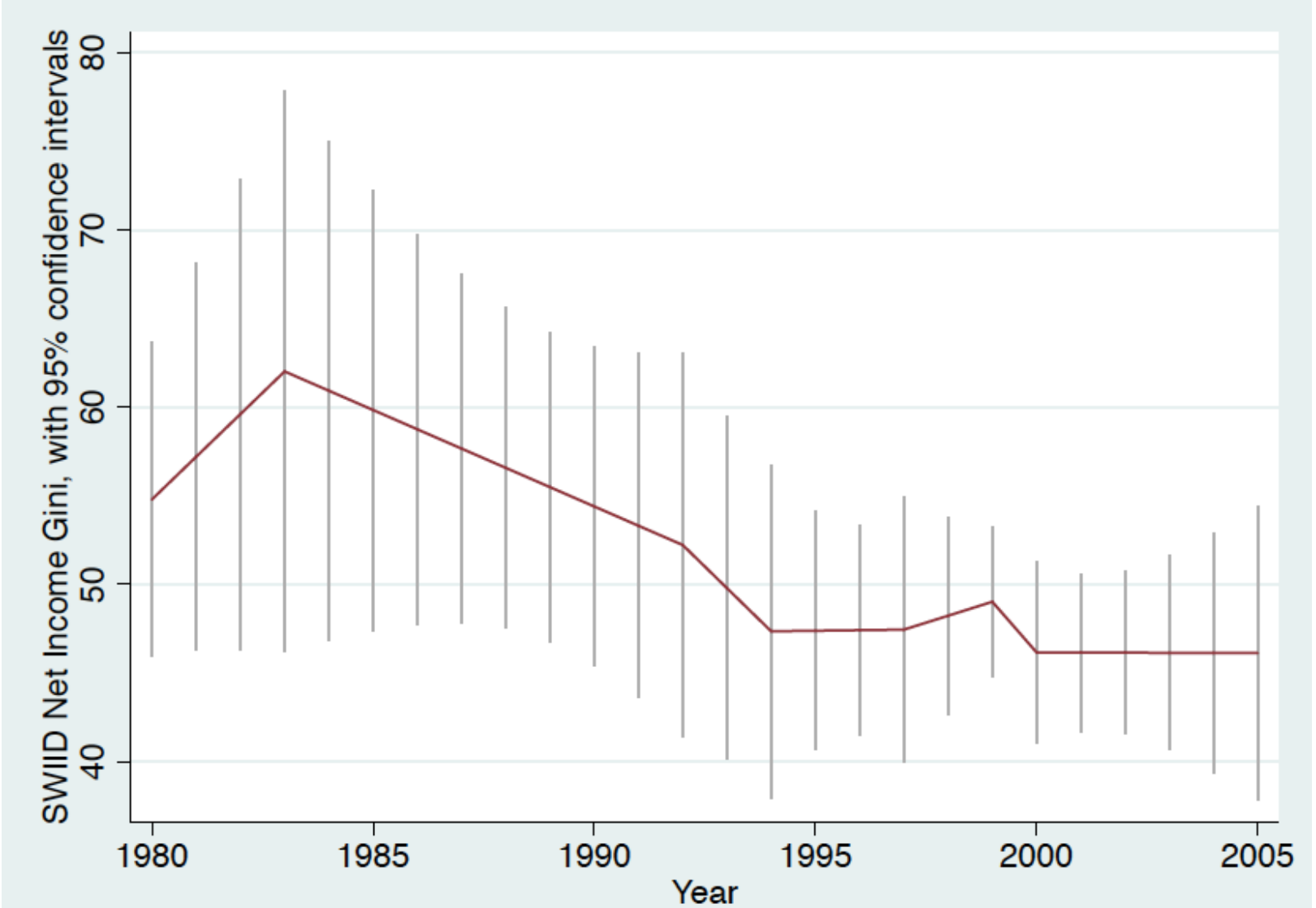

0.80
0.70
0.60
0.50
0.40
0.30
0.20
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
WDI Gini
ATG Gini
SWIID NM Gini
WIID (Average)

Jamaica


0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
WDI Gini
ATG Gini
SWIID NM Gini
WIID (Average)


Kenya
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2007
WDI Gini
ATG Gini
SWIID NM Gini
WIID (Average)

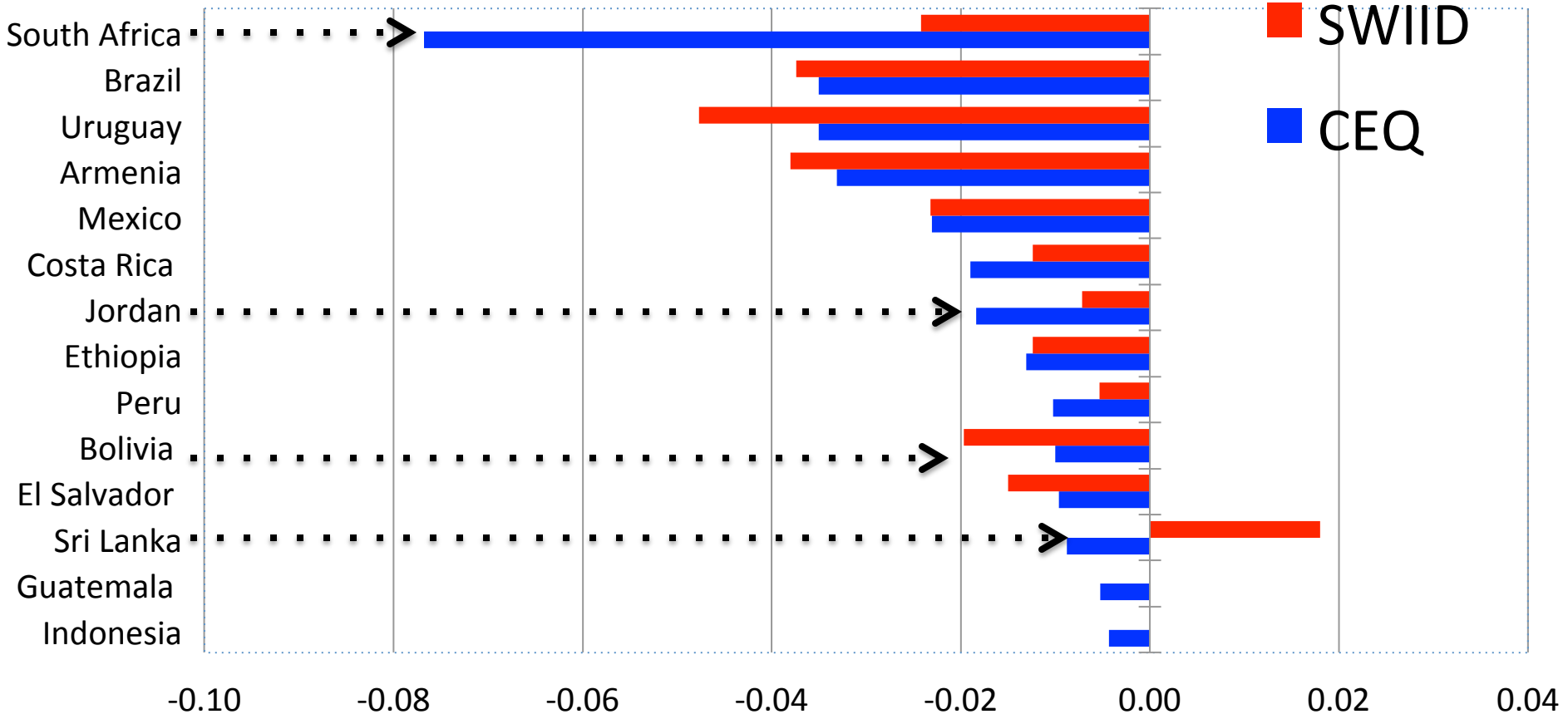
Kenya: Confidence Interval in SWIID



Measuring the Redistributive Effect

- Gini before taxes and transfers vs. Gini after taxes and transfers
- CEQ (Commitment to Equity project): detailed fiscal incidence analysis of taxes and transfers to measure their impact on inequality and poverty
- CEQ vs. SWIID: FOUR out of FOURTEEN cases in which difference in the redistributive effect is significant
  - In three SWIID was lower and higher in one

Change in Gini: Disposable vs. Market
SWIID vs. CEQ
(in GINI points)



CEQ (Commitment to Equity): www.commitmentoequity.org-- Soutce Lustig (2014)

Criteria for the Assessment

- Accessibility and User-friendliness
- Quality of Documentation
- Reliability/Accuracy of Reported Indicators
- Transparency and Replicability

Thank you