

PATHWAY 2: DEVELOPING COMPOSITE INDICATORS

Christopher Gandrud

SG1022, City University London

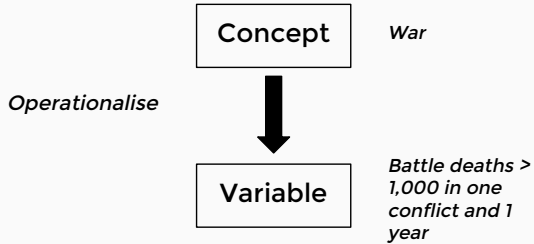
AIMS

- What are composite indicators and who uses them?
- Defining the construct and what can be observed.
- Making Composite Indicators
 - What data to include?
 - How to normalise the variables?
 - How to weight the variables?
 - Assessing validity (introduction)
- Pros, cons, and pitfalls
- Survey examples

WHAT ARE COMPOSITE INDICATORS

Last semester, we learned about concepts and variables.

- *Concept*: A phenomenon (e.g. poverty, democracy, human development, trust in institutions) we are interested in studying.
- *Variable*: Observable characteristic of a unit (e.g. person, city, country) that operationalises the concept.



WHEN ONE VARIABLE IS NOT ENOUGH

Frequently in social science we are interested in complex concepts that cannot be operationalised with one variable.

Instead, they likely involve some combination of variables.

For example, what is democracy?

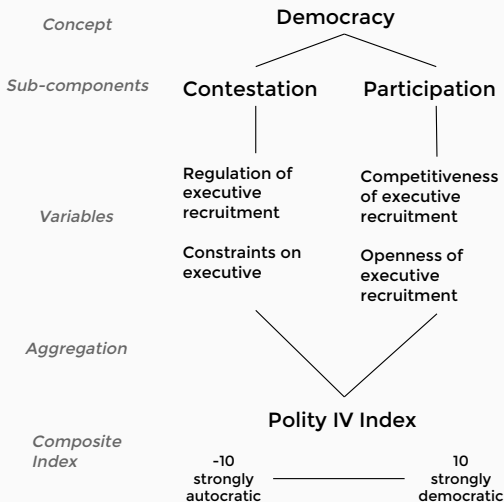
Conceptualisation: Dahl (1971) argued that democracy has two core attributes:

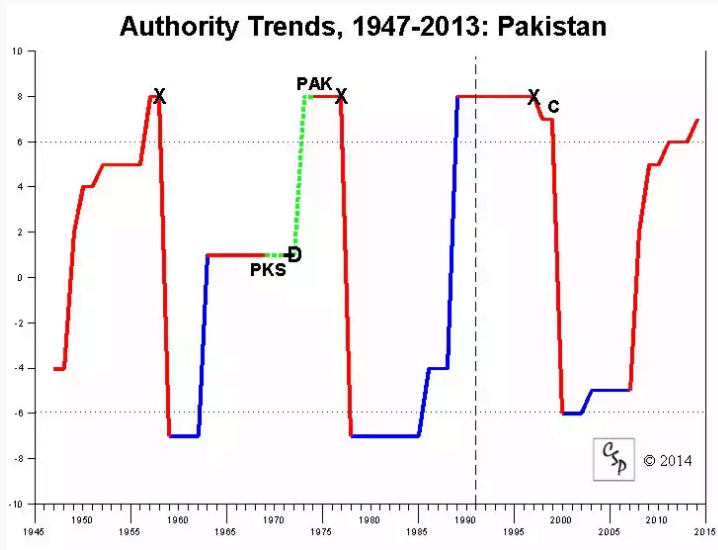
- *contestation*: competitive elections to choose leaders,
- *participation*: inclusive rules for and rates of participation.

Note: not just elections/no elections or universal suffrage/limited suffrage. So ...

...we need a composite indicator to operationalise Dahl's conceptualisation of democracy.

CREATING THE POLITY IV DEMOCRACY INDEX





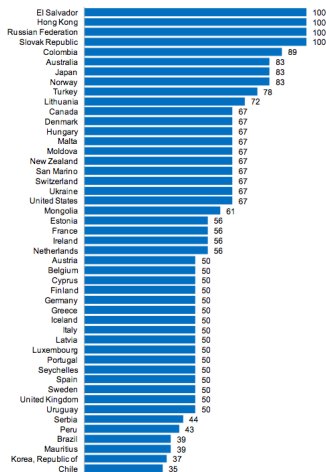
Why should we care about measuring concepts well?

If we don't have good measures of our concepts, we can't know how one thing effects another (relationships *between* concepts) and how to improve the social world.

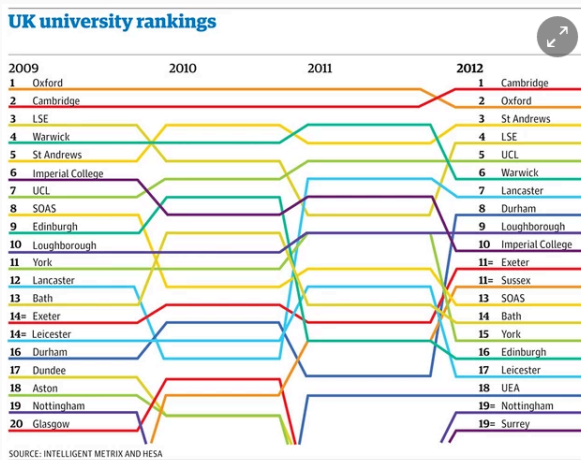
Composite indicators are a popular tool used by governments, international organisations, NGOs, and public policy advocacy groups.

IMF FISCAL TRANSPARENCY (GFS) INDEX

Figure 2. Average Comprehensiveness of GFS, 2011–13: Top Quartile



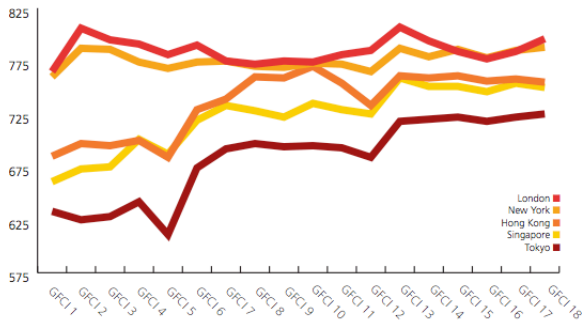
UNIVERSITY RANKINGS



(<http://www.theguardian.com/news/datablog/2011/may/17/university-guide-2012-data-guardian>)

GLOBAL FINANCIAL CENTRES INDEX (2015)

Chart 4 | Top Five Centres GFCI Ratings Over Time



New York and London vie for crown of world's top financial centre

Financial centres face competition from Asia and increased regulation

Financial Times (1 Oct 2014)

STEPS TO CREATING COMPOSITE INDICATORS

1. Theoretical framework
2. Data selection
3. Address missing data
4. Multivariate analysis (note: we don't really cover this in this course)
5. Normalisation
6. Weighting and aggregation
7. Validation

Modified from OECD (2008, 20-21)

All of these steps should be clearly documented in detail.

THEORETICAL FRAMEWORK

“What is badly defined is likely to be badly measured.” (OECD 2008, 22)

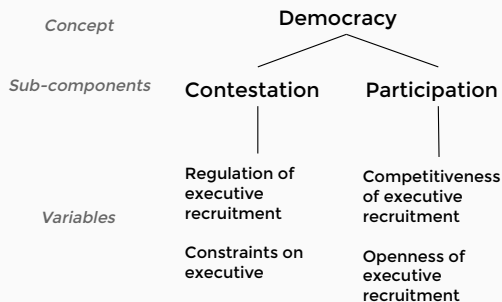
Concept definitions should:

- be clear and parsimonious,
- refer to a theoretical framework,
- select a relevant unit of analysis,
- link various sub-groups and the underlying indicator.

Sub-components:

- should be *statistically independent* of each other (want as few components as possible, don't want to double up),
- existing *linkages* between them should be *described theoretically and empirically* as much as possible.

CONCEPTUALISING DEMOCRACY



DATA SELECTION

A composite indicator is only as good as its parts.

Data on the indicators' component should be selected based on at least their:

- relevance to the theoretical framework,
- analytical soundness,
- timeliness,
- accessibility.

Gathering, cleaning, and merging the component variables is not trivial.

Has important substantive implications (e.g. error can bias your indicators) and can be time consuming.

So, needs to be well documented and reproducible.

MISSING DATA

To aggregate your components into a composite indicator you need a complete data set, with no missing values.

For example, imagine we are going to make a component indicator with variable_1, variable_2, and variable_3. A complete data set would look like:

country	variable_1	variable_2	variable_3
Algeria	1	10034	0
Cambodia	0	30020	0
Zambia	0	50302	1

However, we often (especially when working with country-level data) have missing data on at least some of our variables.

For example:

country	variable_1	variable_2	variable_3
Algeria	NA	10034	0
Cambodia	0	30020	0
Zambia	0	NA	1

Note: NA in R means missing value.

You should always assess how much missing data you have and consider why the data is missing.

Data can be:

- *Missing at random*: why the variable value is missing has nothing to do with the variable—random.
- *Not missing at random*: the variable value is missing because of some reason related to the variable.
 - E.g.: Low income countries don't have the money to pay for enough staff to gather GDP data.

Not missing at random is complex to deal with and requires explicit modeling of the missing process.

We do not cover this modeling process in this course.

However, if you suspect your data is not missing at random, you should discuss this including the reasons why you believe this.

If you have missing data, to get complete cases you can:

- drop incomplete cases,
- single impute data (e.g. replacing the missing value with the variable's mean/median/mode),
- multiple impute data (not covered in this course).

ANALYSIS OF THE COMPONENT STRUCTURE

“Analysing the underlying structure of the data is still an art.”

(OECD 2008, 25)

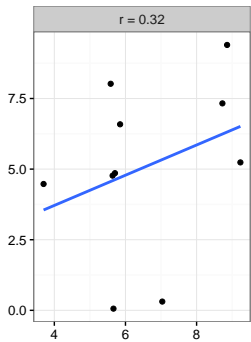
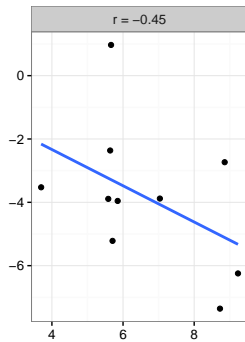
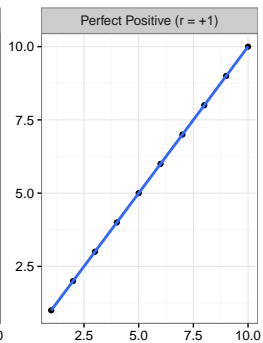
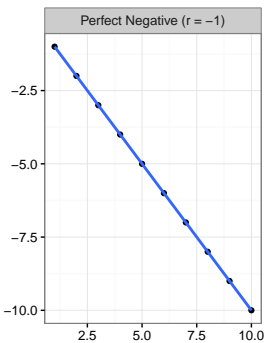
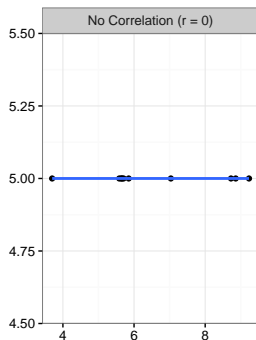
Understanding the underlying structure is important for informing weighting and aggregation decisions.

Note: important multivariate analysis tools–e.g. principal component analysis, cluster analysis–are beyond the scope of this class.

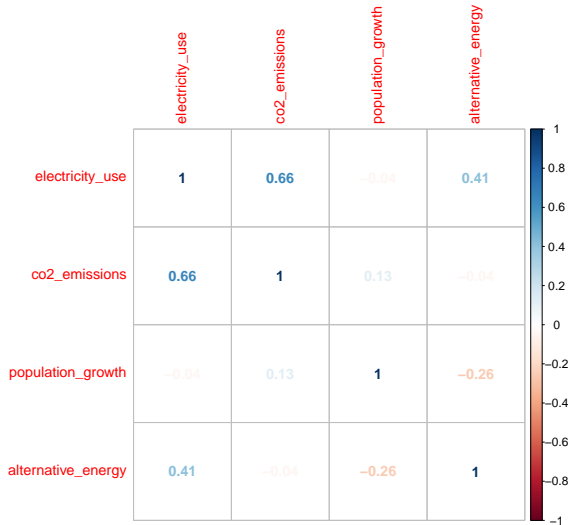
A simple way to examine the relationships between your sub-components is to create a correlation matrix.

Correlation: a mutual relationship between two variables.

Correlation coefficient: a number ranging from -1 to 1 calculated to represent the linear correlation between two variables. Usually denoted r .



CORRELATION MATRIX



NORMALISE VARIABLES

Observable variables are often on different scales as they can have different measurement units.

For example, *life expectancy at birth* is in years ranging from 0 to > 100 and *GNI per capita* is in US dollars starting from > 300.

Obviously, adding these two variables together would weight GNI more than life expectancy.

Some notation:

$x_{i,t}$: Observed variable value for unit i (e.g. country) at time t (e.g. year)

$l_{i,t}$ Normalised value of the variable value for unit i at time t

WEIGHTING AND AGGREGATION

Once you have your normalised variables ($I_{c,t}$), then you need to consider how to combine them.

Things to consider:

- **Weighting:** how important are each individual variables to the composite?
- What scale do you want the indicator to be on?

“No uniformly agreed methodology exists to weight individual indicators before aggregating them into a composite indicator.”

But generally, “what matters more ...weighs more.”

<https://composite-indicators.jrc.ec.europa.eu/?q=content/step-6-weighting>

If you simply add all of the variables together, you are implicitly assuming that they have an equal weight of 1.

$$Cl_{c,t} = \sum I_{c,t} * 1 \quad (1)$$

Sometimes this makes sense: e.g. economic activity is often measured in the same currency.

Sometimes the concept we are measuring might be discrete. E.g. you are in a financial crisis or not in a financial crisis.

So, you might set a **threshold**, a point past which a unit goes from having the characteristic to not having the characteristics.

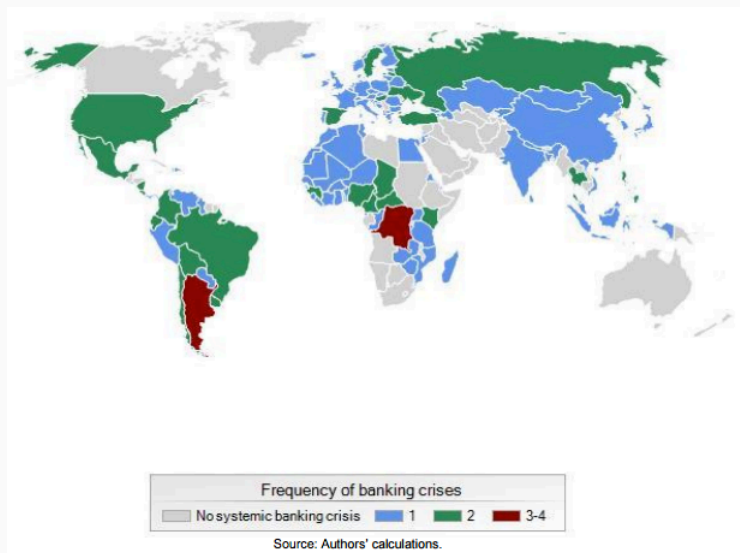
Laeven and Valencia (2013) determine a country has crossed a financial crisis threshold if:

- There is 'significant distress' in a country's financial system.

and

- At least three of six policy responses are used (e.g. bank holidays, bank nationalisations).

LAEVEN & VALENCIA BANKING CRISES (1970-2011)



WHAT ARE WE MEASURING?

Once you have a composite indicator, your work is far from done.

You need to conduct numerous tests to determine if your indicator is a valid measure of the concept you are trying to measure.

There is likely **no perfect composite indicator** for any given concept.

Transparency—conceptualisation, data gathering/cleaning, construction, validation—is crucial for others to be able to understand and evaluate your indicators.

PROS AND CONS

However, we need to remember that our indicators are **estimates** of what we want to measure.

We are **uncertain** about how well our indicators capture reality.

Uncertainty can be caused by at least:

- Error in our construct (i.e. including or omitting important variables)
- Measurement error in our raw data $x_{c,t}$
- Error in our weighting.

Composite indicators are popularly used to rank units (e.g. cities, countries).

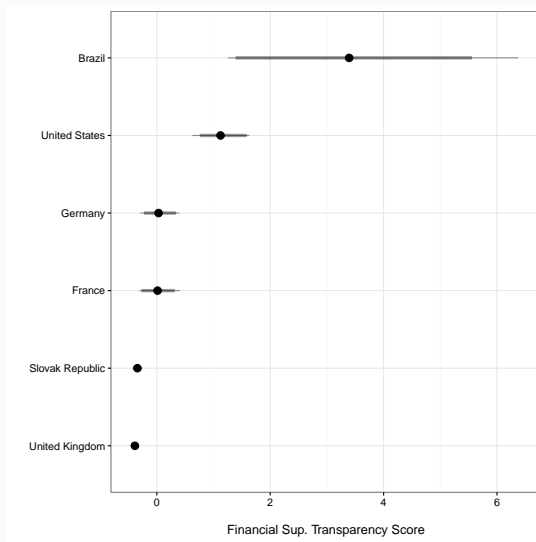
However, this can be highly misleading.

For example, Copelovitch, Gandrud, and Hallerberg (2015) created a measure of financial supervisory transparency.

Table: Financial Regulatory Transparency Index Ranking (2011)

Country	Rank
Brazil	1
United States	7
Germany	23
France	25
Slovak Republic	48
UK	61

INDICATOR WITH ABSOLUTE SCORE & UNCERTAINTY



In this course, we won't be covering more advanced ways to quantify your uncertainty about your composite indicators.

However, when you should be careful and honest when you compare units by ranking their composite scores.

Be honest about what you don't know.

PROS AND CONS OF COMPOSITE INDICATORS

Pros	Cons
Summarise complex multi-dimensional concepts	May be misleading if poorly constructed/non-transparent/interpreted
Support policy-making	May lead to simplistic policy conclusions or missed to support particular predefined policy goal
Facilitate communication with the public.	May disguise problems in one dimension if construction is not transparent. May exaggerate differences between units if uncertainty is not acknowledged.

Partially from OECD (2008)