

PS0700

Nuts and Bolts of Political ‘Science’: Measurement

Political Science Research Methods

Professor Steven Finkel

Fall Semester 2022

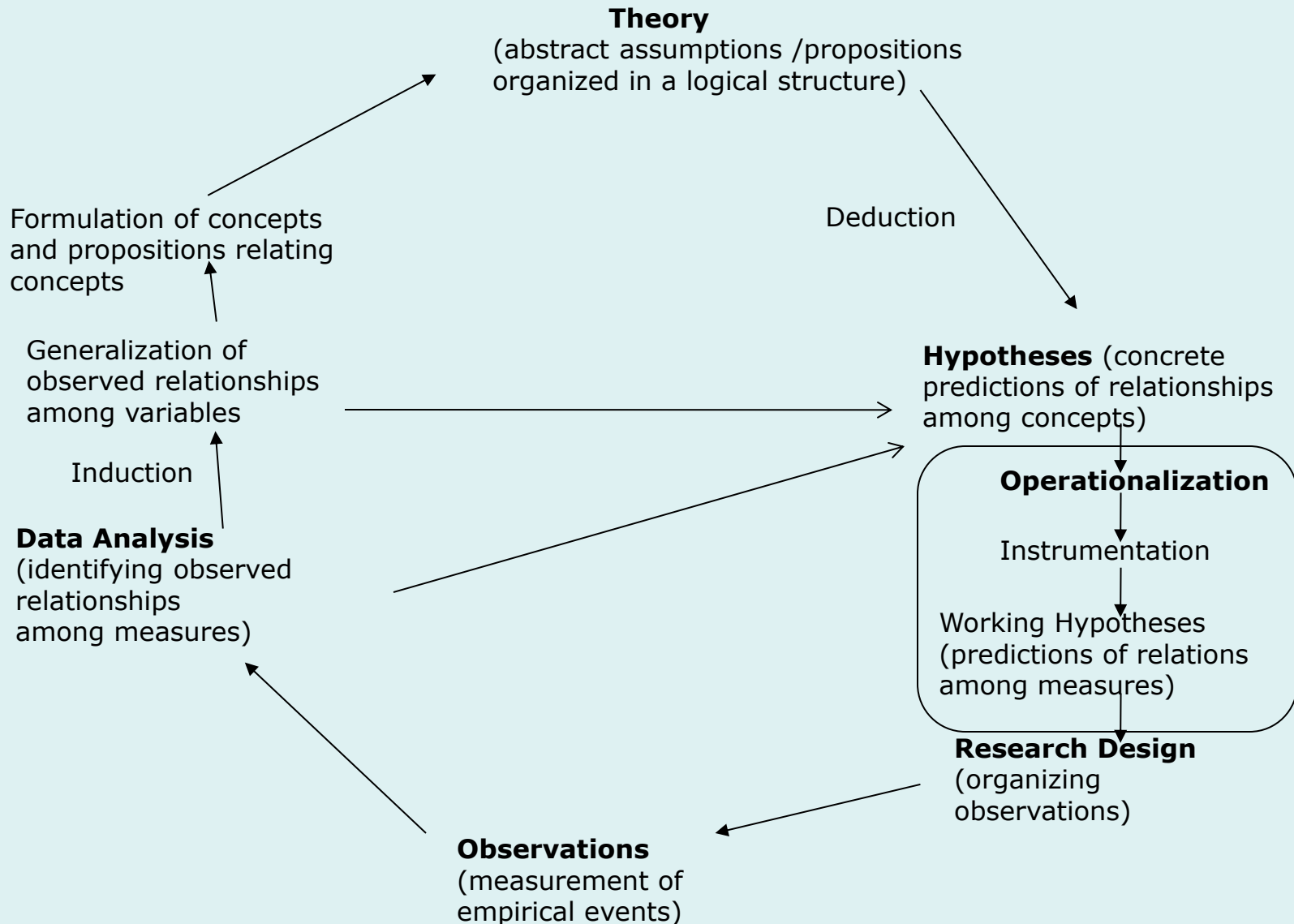
Week 3a



Goals for the Session

- Provide overview of measurement process in political science research
- Discuss concept of “measurement error” and ways to assess the accuracy of measurements

A Model of the Research Process



Measurement in Political Science

- Once a hypothesis has been adopted, next step is to move towards testing it with observable data
- To do this, we need first to “operationalize” the variables in our hypothesis, i.e., *define the variables in such a way that they can be observed empirically and measured*
- *Measurement* itself is done only after operational definitions have been laid out. We can conceive of “measurement” as the process by which every unit or case is given a value on each variable according to the operational definition of that variable
- Idea is to construct operational definitions that come as close as possible to the “meaning” of the variable, and that will therefore allow measurements to be made that accurately and precisely reflect the variables of interest

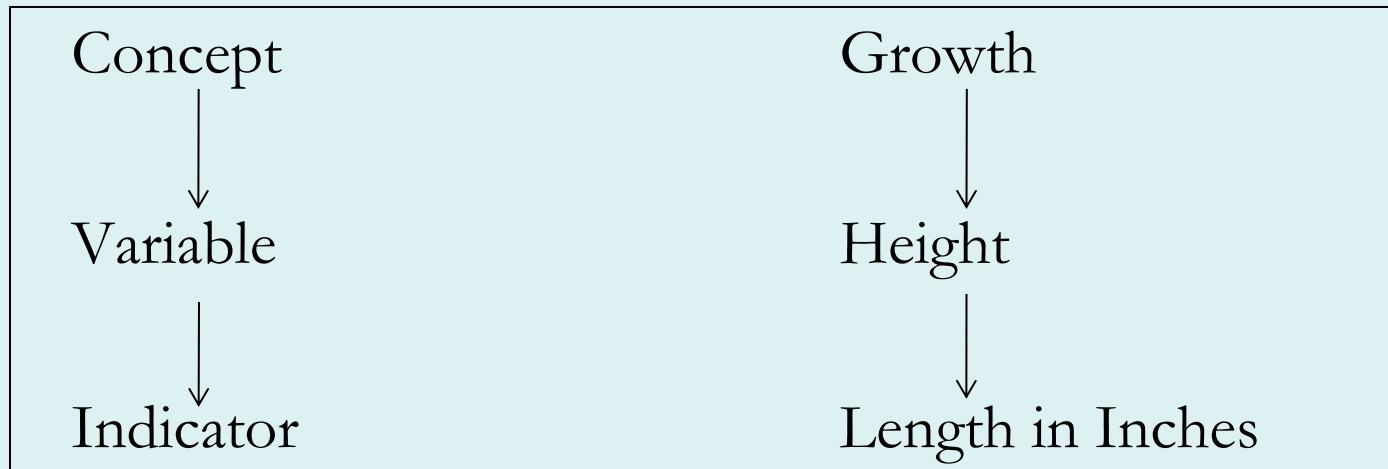
- Further goal: establish measurement procedures that can be replicated by others, so that if they followed the procedures, they would yield the same results
- Measurement is a difficult process in political science because the concepts we are interested in are often multi-dimensional, and different scholars from different theoretical traditions define and use those concepts differently
- **Many** controversies exist in the political science literature about the proper *measurement* of concepts and variables, and how different measurement strategies yield different substantive results
- Example: “Political Tolerance” controversy (Sullivan *et al.* article from syllabus)

From Concepts to Variables to Measurement

- In important way, measurement starts very early in research process: we developed hypotheses in the first place by generating the “*variables*” that correspond to the broader “concepts” in which we are interested
- Example from natural sciences: we want to test the effects of a chemical fertilizer versus natural nutrients on “growth” of corn
 - Concept is “growth”
 - What variable could correspond to “growth”? How can we “operationalize” this concept?
 - We say “height of cornstalk”, so that becomes the *variable* in our hypothesis
 - But: many other ways we might have chosen to operationalize the concept “growth”, for example, weight of the corn, the diameter of the stalks, width of the leaves, etc.

- Now, we need to figure out how we will assign concrete measurements for each unit on the variable “height”, i.e. we have to *operationalize* the variable “height”. How will we come up with the actual measures, or what we call the “indicator(s)”?
- Several possibilities:
 - Ask people to come to cornfield and assess each stalk as “tall” or “short”. Problems of subjectivity, vision, angle of observation, etc., so probably we will reject this method
 - Measure the height of the corn stalks in inches with the aid of a tape measure
- If we agree that the latter strategy will provide a more objective assessment, then we have “operationalized” the variable “height” as “length in inches”, and that will be our “indicator” of the variable
- We then observe each cornstalk and “measure” it, or assign values to it, based on our operational definition

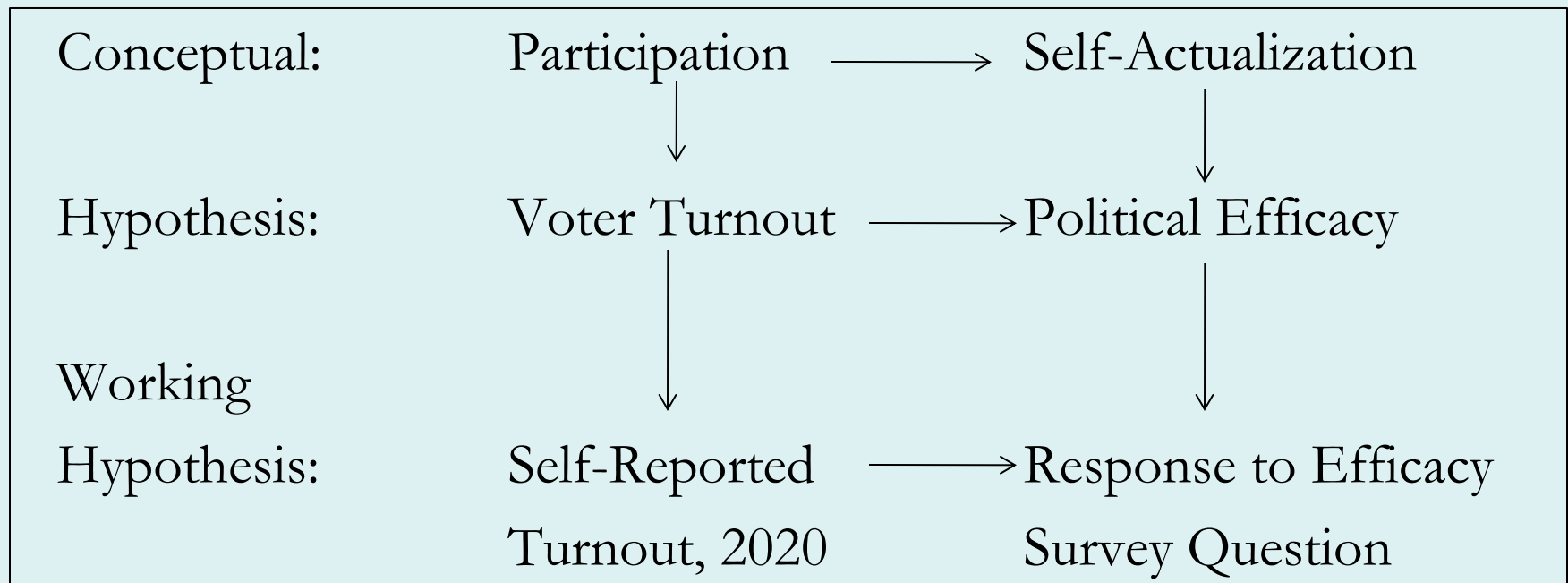
- So the measurement model looks like:



- Important --- we don't actually test the relationship between *concepts*, or even the relationship between *variables*, we test the relationship between *operationalized indicators* of variables
- Our empirical tests, then, are only as good as the operationalizations and indicators we come up with!
- We inevitably lose information and meaning in the operationalization process. We try to minimize this loss!

Political Science Example

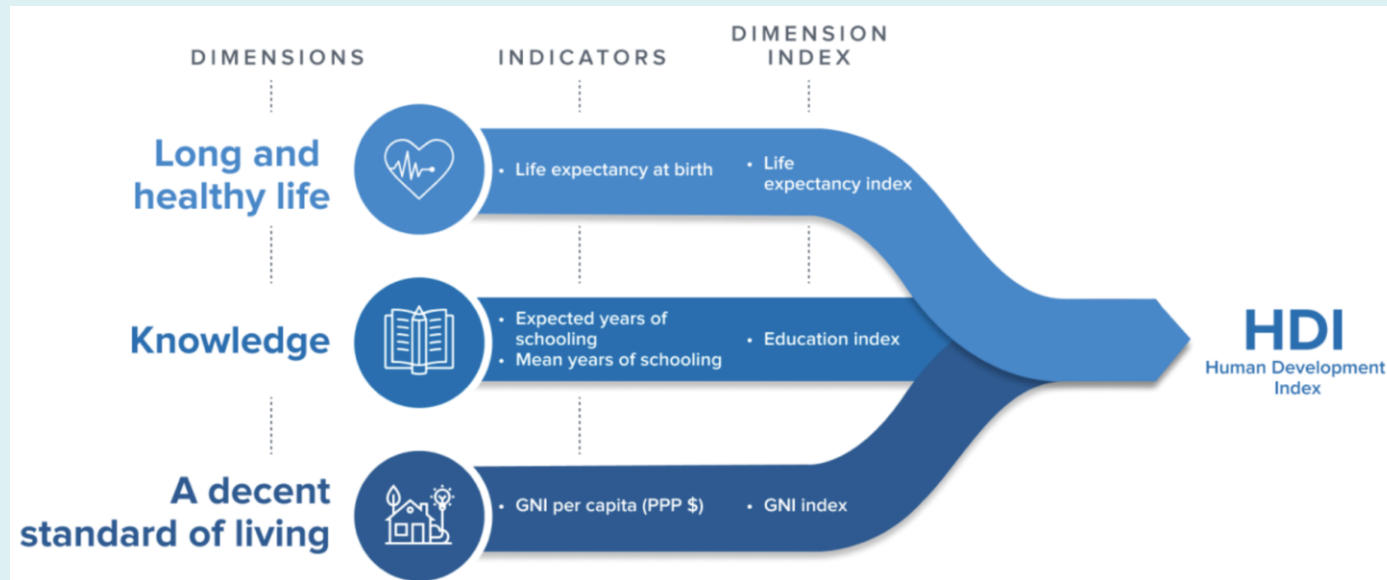
- Conceptual Relationship: Political Participation has a positive “self-actualization” or “developmental” effect on individuals
- Variables:
 - Voter Turnout
 - “Political Efficacy” (sense of competence in politics)
- Operationalized Indicators
 - Self-reported turnout in past presidential election
 - Response to question “People like me have no say in what the government does” (this is a real question in the National Election Studies conducted by the University of Michigan
www.electionstudies.org)



- The *horizontal arrows* represent our causal hypotheses or proposed regularities; the *vertical arrows* represent the operationalization process that results in a “working hypothesis” with concrete indicators that will be used in subsequent empirical tests
- There will ***always*** be “measurement error” in this process, and ***all*** political science research can be improved in terms of measurement
- If our indicators are poor, we will not know if the relationships we find in the empirical tests are an accurate reflection of the “true” causal relationship between the variables or the concepts. That is why measurement is so important in political science!!!!

Application: The UNDP Human Development Index

<http://hdr.undp.org/en/content/human-development-index-hdi>



Data sources

- Life expectancy at birth: UNDESA (2022a).
- Expected years of schooling: CEDLAS and World Bank (2022), ICF Macro Demographic and Health Surveys (various years), UNESCO Institute for Statistics (2022) and United Nations Children's Fund (UNICEF) Multiple Indicator Cluster Surveys (various years).
- Mean years of schooling: Barro and Lee (2018), ICF Macro Demographic and Health Surveys (various years), OECD (2022), UNESCO Institute for Statistics (2022) and UNICEF Multiple Indicator Cluster Surveys (various years).
- GNI per capita: IMF (2022), UNDESA (2022b), United Nations Statistics Division (2022) and World Bank (2022).

Dimension	Indicator	Minimum	Maximum
Health	Life expectancy (years)	20	85
Education	Expected years of schooling (years)	0	18
	Mean years of schooling (years)	0	15
Standard of living	Gross national income per capita (2011 PPP \$)	100	75,000

$$\text{Dimension index} = \frac{\text{actual value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}} \quad (1)$$

ack

Table 1. Human Development Index and its components

		SDG3		SDG4.3		SDG4.4		SDG8.5			
		Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling		Mean years of schooling		Gross national income (GNI) per capita		GNI per capita rank minus HDI rank	HDI rank
DI rank	Country	Value	(years)	(years)		(years)		(2017 PPP \$)			
		2021	2021	2021	a	2021	a	2021		2021	2020
	VERY HIGH HUMAN DEVELOPMENT										
1	Switzerland	0.962	84.0	16.5		13.9		66,933		5	3
2	Norway	0.961	83.2	18.2	c	13.0		64,660		6	1
3	Iceland	0.959	82.7	19.2	c	13.8		55,782		11	2
4	Hong Kong, China (SAR)	0.952	85.5	17.3	d	12.2		62,607		6	4
5	Australia	0.951	84.5	21.1	c	12.7		49,238		18	5
6	Denmark	0.948	81.4	18.7	c	13.0		60,365		6	5
7	Sweden	0.947	83.0	19.4	c	12.6		54,489		9	9
8	Ireland	0.945	82.0	18.9	c	11.6	e	76,169	f	-3	8
9	Germany	0.942	80.6	17.0		14.1	e	54,534		6	7
10	Netherlands	0.941	81.7	18.7	c,e	12.6		55,979		3	10
11	Finland	0.940	82.0	19.1	c	12.9		49,452		11	12
12	Singapore	0.939	82.8	16.5		11.9		90,919	f	-10	10
13	Belgium	0.937	81.9	19.6	c	12.4		52,293		7	16
13	New Zealand	0.937	82.5	20.3	c	12.9		44,057		16	13
15	Canada	0.936	82.7	16.4		13.8	e	46,808		9	15
16	Liechtenstein	0.935	83.3	15.2		12.5	g	146,830	f,h	-15	14
17	Luxembourg	0.930	82.6	14.4		13.0	i	84,649	f	-13	17
18	United Kingdom	0.929	80.7	17.3		13.4		45,225		9	17
19	Japan	0.925	84.8	15.2	e	13.4		42,274		12	19
19	Korea (Republic of)	0.925	83.7	16.5		12.5	e	44,501		9	20
21	United States	0.921	77.2	16.3		13.7		64,765		-14	21
22	Israel	0.919	82.3	16.1		13.3	e	41,524		10	22
23	Malta	0.918	83.8	16.8		12.2		38,884		12	26
23	Slovenia	0.918	80.7	17.7		12.8		39,746		10	23
25	Austria	0.916	81.6	16.0		12.3		53,619		-8	23
26	United Arab Emirates	0.911	78.7	15.7		12.7		62,574		-15	25
27	Spain	0.905	83.0	17.9		10.6		38,354		10	27
28	France	0.903	82.5	15.8		11.6		45,937		-2	28
29	Cyprus	0.896	81.2	15.6		12.4		38,188		9	29
30	Italy	0.895	82.9	16.2		10.7		42,840		0	32
34	France	0.888	77.4	15.0		10.5		33,646		0	33

Measurement Error: The Enemy

- How do we evaluate the adequacy of indicators? We want the indicators to reflect accurately the variables they are supposed to represent. That is, we want whatever value we assign to a unit on the given indicator to reflect that unit's "TRUE" value on the underlying variable, not "error"
- In simple equation form:
$$X = T + \varepsilon$$
where X is the value assigned on the indicator, T is the "True Score" on the variable of interest, and ε is the error in measurement
- So every measurement contains some combination of the "True" score plus some error, which we call "measurement error" and which we try to minimize in practice

Types of Measurement Error

- We classify measurement error as either “*systematic*” or “*random*”
 - *Systematic errors* are errors that make *every* observation either higher or lower than it really should be, i.e., it leads to all units being assigned different values on X than they really are on T. If this happens, we say that the X is an *invalid* indicator or measure of T.
 - *Random errors* are errors that affect every unit differently, so that some units score higher on X than they truly are on T, and some units lower. These kinds of errors result in X being an *unreliable* indicator or measure of T.

- Simply stated, an indicator is *valid*, or without systematic error, if it measures what it is supposed to measure. That is, X really does get at the T that you want it to get at. If it doesn't, there are systematic biases in the measure, and X doesn't really measure what you want it to measure
- An indicator is *reliable*, or without random error, if it generates a *consistent* response. If you took the same measurement 10 times for a given indicator, would you assign the same value each time, or would some random factors lead to assigning higher or lower values at each measurement observation? If consistent, we say the measure is “reliable”; if inconsistent, we say it is “unreliable”
- Measures can be *reliable* but not *valid* – (consistently bad!)
- Measures can be *valid* but not *reliable* – (inconsistently good!)
- This last statement contradicts the figure in the text on p.116 (This is a subtle point but I am correct 😊)

Examples

- We think that income will affect vote choice
 - Possible indicators of income:
 - Self-reported yearly income
 - IRS W-2 form (reported income on tax return)
 - Neighbor's evaluation of how wealthy someone is
 - Can you think of *validity* problems (systematic error) and *reliability* problems (random error) with these indicators? Are these measures likely to get at the person's "income" accurately and consistently?
- Other examples (from Class Web Site, for Recitation Discussion)
 - Measuring Exposure to Second-hand Smoke
 - Measuring Racial Attitudes
 - Measuring Age (!)
 - Measuring Political Tolerance (Sullivan *et al* 1979; assigned reading)
 - Measuring Democratic/Autocratic Nature of Political Regimes (Luhmann *et al.* (2018; assigned reading)

Assessing Indicator Validity

- How do we determine if an indicator (or set of indicators) is *valid* or not? Extremely difficult!!
- Most common methods:
 - **“Face validity”** and **“content validity”**– does the measure on its face appear to tap the relevant variable accurately (*“face validity”*)? Does it tap all the dimensions of the variable (*“content validity”*)? This is a judgment based largely on the common sense of informed experts observers (like us!) As such, it is not terribly “scientific” but almost always indicators are defended at least partially on face and/or content validity grounds
 - **“Construct validity”** – does the indicator perform well in predicting what it is supposed to predict? Is it related to variables that are known to be related to T? If an indicator, e.g., of social class *did not* correlate with education, we would say it did not have construct validity. *Problem: What if social class really isn't related to education in a particular setting?*

- **“internal validity”, or “inter-item association”** --- if you use multiple indicators to measure a variable, are the indicators highly correlated with one another? If there are high correlations between all of the indicators designed to measure the same variable, then chances are that those indicators are valid ones. If there is a high correlation between several of the indicators, but one of the indicators is not correlated with the others, then that indicator may be invalid

Example: Measuring teacher quality with three indicators:

Student Evaluations

Peer Evaluations by other Faculty

Class Enrollment Size

If these indicators are highly intercorrelated, we have evidence of valid indicators of the variable “teacher quality”. It is likely, though, that enrollment size will not be correlated with the other two, so it is probably not valid

Assessing Indicator Reliability

- All indicators are unreliable to some degree because of random factors influencing the measurement process
 - Respondent's mood
 - Interpretation of questions by respondents
 - Context and administration of the instrument used to obtain the measurement
 - Scaling of responses (a person who is truly in the middle of “strongly agree” and “agree” will say one thing one day, another thing the next)
 - Units like cities or countries may differ in the methods they use to collect and report data on crime, expenditures, participation, etc.
- This is why you can have unreliable measures that are still valid – there is intrinsic unreliability in the measurement process. But after a point – unreliable measures become useless to analyze

- General recommendation for improving indicator reliability:
Use multiple indicators!!!
 - Random measurement error in multiple indicators will tend to cancel out, and thus, even if each indicator is not perfectly reliable, a combined scale from **all** of the indicators will tend to converge on the true value of the variable T
 - There are many reliability tests available when multiple indicators are present (e.g. “split-halves”, “Cronbach’s alpha”, “inter-item correlations”, etc.)
- Multiple indicators also can help with validity issues
 - Multiple indicators have a greater likelihood of covering all the dimensions of variables that the researcher wants to get at
- Be especially skeptical of research with single indicators of the variables!

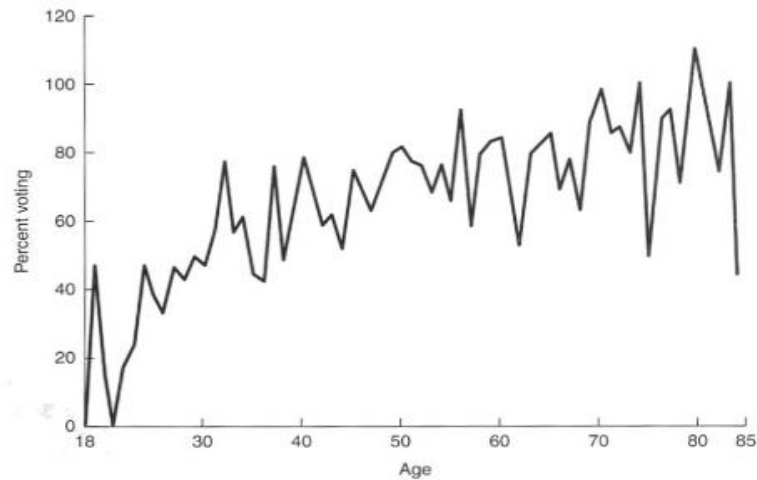
Levels of Measurement

- The next step is the *measurement* of the indicators for each unit, where we assign each unit a particular numerical value corresponding to “how much” of that indicator a given unit possesses. How precise that measurement is, what those numbers mean, and how we can manipulate and interpret them, depends in turn on the type of variable that we are measuring, on what is referred to as the variable’s “*level of measurement*”
- Four Types of Measures:
 - **Nominal:** categories that have no intrinsic ordering or ranking
 - **Ordinal:** ordered or ranking of units is possible, but there is no fixed or meaningful distance between categories
 - **Interval:** ordered with equal distance between categories
 - **Ratio:** ordered with equal distance between categories *and* a meaningful zero point

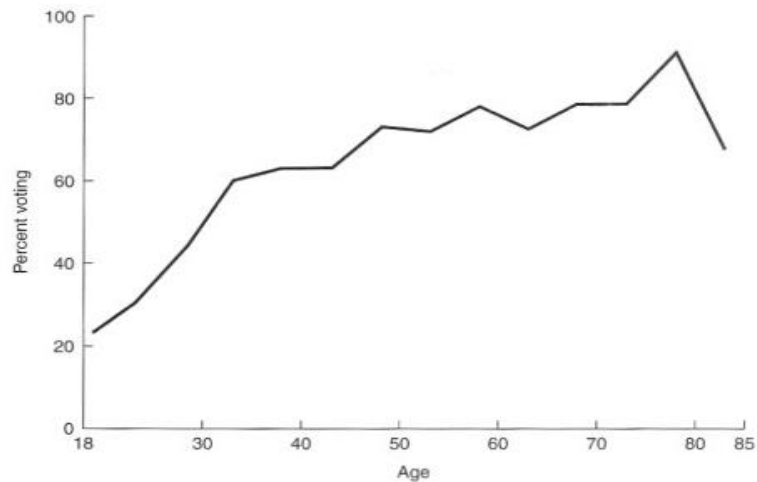
Examples

- Nominal Measures
 - Religion (Protestant, Catholic, Jewish, Muslim, Other)
 - Nationality
 - Note: Nominal schemes must be *mutually exclusive* (units can only be assigned to one category) and *collectively exhaustive* (all units can be assigned to *some* category)
- Ordinal Measures
 - Social Class (lower, middle, upper)
 - Prestige of Political Science Department (low, medium, high)
- Interval Measures
 - Temperature in Fahrenheit
- Ratio Measures
 - Age
 - Unemployment rate
 - Budgetary expenditures

- Why is the level of measurement important?
 - Higher Level of Precision is Often Desirable from Research (and from Policy) Standpoint.
 - We usually want the highest level of measurement possible so that we waste as little information as possible. We want to say, for example, that if we increase the police budget by 1000 dollars per year, we will reduce the crime rate by 5 crimes per 1000 residents, as opposed to saying we will reduce crime by a “medium” (compared to a “small” or “large”) amount, or simply we will “reduce” crime (compared to “not reducing” crime). This gives maximum precision to our our predictions and gives us greater understanding of the causal processes
 - Not always the case, though, as sometimes interval or ratio gives you too much information, especially for visual displays like graphs. So you should gather the data in interval or ratio form if possible, but not necessarily use it in this form in all of your analyses (see next slide)



(a) Age and participation in 2002 election: age measured by years





(b) Age and participation in 2002 election: age measured by half-decades

Figure 5.3

An example of the effect of grouping data on interpretation

SOURCE: W. Philips Shively, *The Craft of Political Research*, 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2004, pp. 59–60. Reprinted with permission.

- Finally, the level of measurement is important because it determines the statistical procedure to use in testing relationships between variables, as we will see in quantitative analysis section of the class

<div>Independent </div> <div>Dependent </div>	Nominal/Ordinal	Interval/Ratio
Nominal/Ordinal	Crosstabulation	Logistic Regression
Interval/Ratio	T-Test	Regression