

Intermediate Statistics: Measurement

Raymond M. Duch

Nuffield College Oxford

January 18, 2011

What we'll talk about today

1 The problem of measurement

What we'll talk about today

- 1 The problem of measurement
- 2 Issues in measuring concepts of interest

What we'll talk about today

- 1 The problem of measurement
- 2 Issues in measuring concepts of interest
- 3 Examples of measurement problems

Outline of the session

- 1 The problem of measurement
- 2 Issues in measuring concepts of interest
- 3 Examples of measurement problems

Remember what a theory is

- As we have said, a theory is a statement (or a question) about the possible causal relationship between two (or more) concepts.

Remember what a theory is

- As we have said, a theory is a statement (or a question) about the possible causal relationship between two (or more) concepts.
- We have been using both the abstract “Does X Cause Y ?” language as well as the more specific “Does cigarette smoking cause heart disease?” language.

How do we evaluate our theories?

- That is, how do we come to a conclusion about whether our theory is likely to be correct?

How do we evaluate our theories?

- That is, how do we come to a conclusion about whether our theory is likely to be correct?
- We need to make empirical observations. In other words, we need to compare our abstract theoretical ideas with reality. (Remember, “empirical” just means “based on observations.” They might be quantitative or in-depth qualitative.)

There's a potential problem here

- We need to be as confident as possible that our concepts in our theory correspond as closely as possible to our empirical observations.

There's a potential problem here

- We need to be as confident as possible that our concepts in our theory correspond as closely as possible to our empirical observations.
- This is called the problem of measurement.

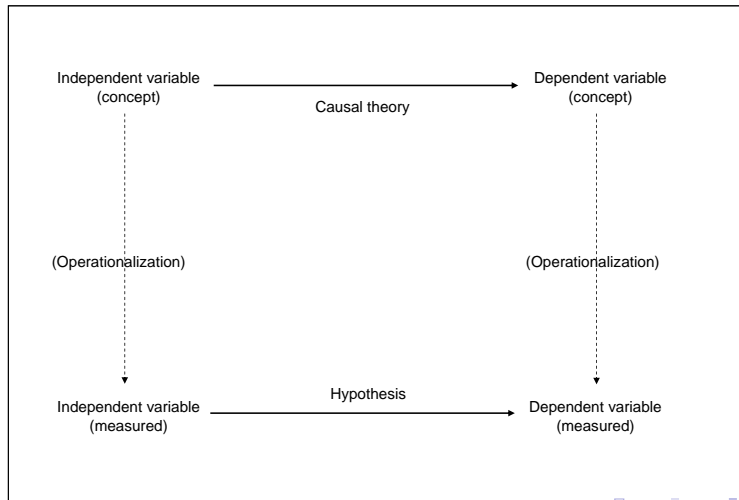
What's the big deal?

- If we want to do a good job evaluating whether X causes Y , then we need to do a precise job measuring both X and Y .

What's the big deal?

- If we want to do a good job evaluating whether X causes Y , then we need to do a precise job measuring both X and Y .
- If we are sloppy in measuring X and Y , then how will we be confident whether or not our assessment of the theory is right?

Recall Figure 1.2



Do you see the disconnect?

- The relationship that we care about most is one we cannot directly observe. We therefore have to rely on potentially imperfect measures of the concepts we care about.

Do you see the disconnect?

- The relationship that we care about most is one we cannot directly observe. We therefore have to rely on potentially imperfect measures of the concepts we care about.
- That means that measuring our concepts with care is one of the most important (and overlooked) parts of social science.

Measurement problems in the social sciences

- **Economics:** Dollars, people

Measurement problems in the social sciences

- **Economics:** Dollars, people
- **Political Science:** ???

Measurement problems in the social sciences

- **Economics:** Dollars, people
- **Political Science:** ???
- **Psychology:** Depression, anxiety, prejudice

Outline of the session

- 1 The problem of measurement
- 2 Issues in measuring concepts of interest**
- 3 Examples of measurement problems

The three issues of measurement

1 Conceptual clarity

The three issues of measurement

- 1 Conceptual clarity
- 2 Reliability

The three issues of measurement

- 1 Conceptual clarity
- 2 Reliability
- 3 Validity

Conceptual clarity

- What is the exact nature of the concept we're trying to measure?

Conceptual clarity

- What is the exact nature of the concept we're trying to measure?
- Example: How should a survey question measure "income"?

Conceptual clarity

- What is the exact nature of the concept we're trying to measure?
- Example: How should a survey question measure "income"?
 - 1 "What is your income?"

Conceptual clarity

- What is the exact nature of the concept we're trying to measure?
- Example: How should a survey question measure "income"?
 - 1 "What is your income?"
 - 2 'What is the total amount of income earned in the most recently completed tax year by you and any other adults in your household, including all sources of income?'

Conceptual clarity

- What is the exact nature of the concept we're trying to measure?
- Example: How should a survey question measure "income"?
 - ① "What is your income?"
 - ② 'What is the total amount of income earned in the most recently completed tax year by you and any other adults in your household, including all sources of income?'
- Example: How should a study measure "poverty"?

Conceptual clarity

- What is the exact nature of the concept we're trying to measure?
- Example: How should a survey question measure "income"?
 - 1 "What is your income?"
 - 2 'What is the total amount of income earned in the most recently completed tax year by you and any other adults in your household, including all sources of income?'
- Example: How should a study measure "poverty"?
- The best measure of a concepts depends on what our theoretical objectives are.

Reliability

- An operational measure of a concept is said to be **reliable** to the extent that it is repeatable or consistent; that is, applying the same measurement rules to the same case or observation will produce identical results.

Reliability

- An operational measure of a concept is said to be **reliable** to the extent that it is repeatable or consistent; that is, applying the same measurement rules to the same case or observation will produce identical results.
- The bathroom scale

Validity

- A **valid** measure accurately represents the concept that it is supposed to measure, while an invalid measure measures something other than what was originally intended.

Validity

- A **valid** measure accurately represents the concept that it is supposed to measure, while an invalid measure measures something other than what was originally intended.
- Example: Measuring prejudice

Validity

- A **valid** measure accurately represents the concept that it is supposed to measure, while an invalid measure measures something other than what was originally intended.
- Example: Measuring prejudice
- Face validity

Validity

- A **valid** measure accurately represents the concept that it is supposed to measure, while an invalid measure measures something other than what was originally intended.
- Example: Measuring prejudice
- Face validity
- Content validity

Validity

- A **valid** measure accurately represents the concept that it is supposed to measure, while an invalid measure measures something other than what was originally intended.
- Example: Measuring prejudice
- Face validity
- Content validity
- Construct validity

Outline of the session

- 1 The problem of measurement
- 2 Issues in measuring concepts of interest
- 3 Examples of measurement problems**

Measuring political tolerance

- Political tolerance and the Blues Brothers

Measuring political tolerance

- Political tolerance and the Blues Brothers
- The Stouffer studies: How did Stouffer measure political tolerance?
Six questions: Should a communist/atheist be allowed to speak/teach/have a book in the library?

Measuring political tolerance

- Political tolerance and the Blues Brothers
- The Stouffer studies: How did Stouffer measure political tolerance?
Six questions: Should a communist/atheist be allowed to speak/teach/have a book in the library?
- What did Sullivan say about these items? Are they valid measures of tolerance? Why or why not?

Measuring political tolerance

- Political tolerance and the Blues Brothers
- The Stouffer studies: How did Stouffer measure political tolerance?
Six questions: Should a communist/atheist be allowed to speak/teach/have a book in the library?
- What did Sullivan say about these items? Are they valid measures of tolerance? Why or why not?
- “Tolerance presumes opposition or disagreement. If there is no reason to oppose, then there is no occasion for one to be tolerant or intolerant” (p. 784) Given this conceptual definition, how did Sullivan measure tolerance?

Measuring democracy

- At the conceptual level, what does it mean to say that Country A is “more democratic” than Country B? Or, a bit differently, what does it mean to say that Country C is becoming “more (or less) democratic” over time?

Measuring democracy

- At the conceptual level, what does it mean to say that Country A is “more democratic” than Country B? Or, a bit differently, what does it mean to say that Country C is becoming “more (or less) democratic” over time?
- The political philosopher Robert Dahl argues that there are two core attributes to a democracy: “contestation” and “participation.”

Measuring democracy

- At the conceptual level, what does it mean to say that Country A is “more democratic” than Country B? Or, a bit differently, what does it mean to say that Country C is becoming “more (or less) democratic” over time?
- The political philosopher Robert Dahl argues that there are two core attributes to a democracy: “contestation” and “participation.”
- Several groups of political scientists have attempted to measure democracy systematically in recent decades. The best-known—though by no means universally accepted—of these is the Polity IV measure. The project measures democracy with annual scores ranging from -10 (strongly autocratic) to +10 (strongly democratic) for every country on earth from 1800 - 2004.

Measuring democracy, part 2

- The Polity IV measure of democracy has four components:

Measuring democracy, part 2

- The Polity IV measure of democracy has four components:
 - 1 Regulation of executive recruitment

Measuring democracy, part 2

- The Polity IV measure of democracy has four components:
 - 1 Regulation of executive recruitment
 - 2 Competitiveness of executive recruitment

Measuring democracy, part 2

- The Polity IV measure of democracy has four components:
 - 1 Regulation of executive recruitment
 - 2 Competitiveness of executive recruitment
 - 3 Openness of executive recruitment

Measuring democracy, part 2

- The Polity IV measure of democracy has four components:
 - 1 Regulation of executive recruitment
 - 2 Competitiveness of executive recruitment
 - 3 Openness of executive recruitment
 - 4 Constraints on chief executive

Measuring democracy, part 3

- For each of these dimensions, experts rate each country on a particular scale. For example, the first criterion, “regulation of executive recruitment,” allows for the following possible values:

Measuring democracy, part 3

- For each of these dimensions, experts rate each country on a particular scale. For example, the first criterion, “regulation of executive recruitment,” allows for the following possible values:
- +3 = regular competition between recognized groups

Measuring democracy, part 3

- For each of these dimensions, experts rate each country on a particular scale. For example, the first criterion, “regulation of executive recruitment,” allows for the following possible values:
- +3 = regular competition between recognized groups
- +2 = transitional competition

Measuring democracy, part 3

- For each of these dimensions, experts rate each country on a particular scale. For example, the first criterion, “regulation of executive recruitment,” allows for the following possible values:
 - +3 = regular competition between recognized groups
 - +2 = transitional competition
 - +1 = factional or restricted patterns of competition

Measuring democracy, part 3

- For each of these dimensions, experts rate each country on a particular scale. For example, the first criterion, “regulation of executive recruitment,” allows for the following possible values:
 - +3 = regular competition between recognized groups
 - +2 = transitional competition
 - +1 = factional or restricted patterns of competition
 - 0 = no competition

Measuring democracy, part 3

- For each of these dimensions, experts rate each country on a particular scale. For example, the first criterion, “regulation of executive recruitment,” allows for the following possible values:
 - +3 = regular competition between recognized groups
 - +2 = transitional competition
 - +1 = factional or restricted patterns of competition
 - 0 = no competition
- Countries that have regular elections between groups that are more than ethnic rivals will have higher scores. By similar procedures, the scholars associated with the project score the other dimensions that comprise their democracy scale.

Criticisms of Polity IV

- On the plus side, the Polity IV measure is rich in historical detail, the coding rules are transparent and clear, and the amount of raw information that goes into a country's score for any given year is impressive.

Criticisms of Polity IV

- On the plus side, the Polity IV measure is rich in historical detail, the coding rules are transparent and clear, and the amount of raw information that goes into a country's score for any given year is impressive.
- But the Polity measure includes only one part of Dahl's definition of democracy. There is a lot of information about what Dahl calls "contestation." But the measure lacks information about a country's level of "participation." This may be understandable, in part, because of the time scope of the study. In 1800, very few countries had broad electoral participation.

Criticisms of Polity IV

- On the plus side, the Polity IV measure is rich in historical detail, the coding rules are transparent and clear, and the amount of raw information that goes into a country's score for any given year is impressive.
- But the Polity measure includes only one part of Dahl's definition of democracy. There is a lot of information about what Dahl calls "contestation." But the measure lacks information about a country's level of "participation." This may be understandable, in part, because of the time scope of the study. In 1800, very few countries had broad electoral participation.
- But if the world is becoming a more democratic place, owing to expansion of suffrage, our measures of democracy ought to incorporate that reality. The Polity IV measure does not fully encompass what it means, conceptually, to be more or less democratic. (This is content validity.)

Cronbach's Alpha: Measure of Scale Reliability

- Measure of internal consistency - how closely related a set of items are as a group

$$\alpha = \frac{N * \bar{c}}{\bar{v} + (N - 1) * \bar{c}} \quad (1)$$

Cronbach's Alpha: Measure of Scale Reliability

- Measure of internal consistency - how closely related a set of items are as a group
- is a function of the number of test item (N), the average inter-correlation among the items (\bar{c}), and the average variance (\bar{v})

$$\alpha = \frac{N * \bar{c}}{\bar{v} + (N - 1) * \bar{c}} \quad (1)$$

Reliability of Trust in Political System Scale

```
. ****TRUST IN THE POLITICAL SYSEM
. *two factor example
.
. global trust trstprl trstlgl trstplc trstplt trstprt trstep trstun
.
. des $trust // 0-10 scale
```

variable name	storage type	display format	value label	variable label
trstprl	byte	%8.0g	LABC	Trust in country's parliament
trstlgl	byte	%8.0g	LABC	Trust in the legal system
trstplc	byte	%8.0g	LABC	Trust in the police
trstplt	byte	%8.0g	LABC	Trust in politicians
trstprt	byte	%8.0g	LABC	Trust in political parties
trstep	byte	%8.0g	LABC	Trust in the European Parliament
trstun	byte	%8.0g	LABC	Trust in the United Nations

Item Correlations

```
. pwcorr $trust [aw=weight], sig
```

	trstprl	trstlgl	trstplc	trstplt	trstprt	trstep	trstun
trstprl	1.0000						
trstlgl	0.6667 0.0000	1.0000					
trstplc	0.5463 0.0000	0.7039 0.0000	1.0000				
trstplt	0.6628 0.0000	0.5449 0.0000	0.4744 0.0000	1.0000			
trstprt	0.6299 0.0000	0.5195 0.0000	0.4343 0.0000	0.8498 0.0000	1.0000		
trstep	0.4695 0.0000	0.4055 0.0000	0.3289 0.0000	0.5217 0.0000	0.5344 0.0000	1.0000	
trstun	0.4133 0.0000	0.3960 0.0000	0.3767 0.0000	0.4795 0.0000	0.4809 0.0000	0.7513 0.0000	1.0000

Cronbach's Alpha

```
. alpha $trust , item
```

```
Test scale = mean(unstandardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem covariance	alpha
trstprl	53362	+	0.8324	0.7536	3.673209	0.8762
trstlgl	53318	+	0.8074	0.7150	3.696197	0.8809
trstplc	54187	+	0.7497	0.6330	3.861636	0.8908
trstplt	53703	+	0.8475	0.7831	3.754276	0.8735
trstprt	53397	+	0.8310	0.7622	3.806338	0.8762
trstep	47932	+	0.7350	0.6330	3.964784	0.8900
trstun	48363	+	0.7325	0.6233	3.93814	0.8915
Test scale					3.814613	0.8979

Factor Analysis

- Estimate underlying latent variables – or scales

Factor Analysis

- Estimate underlying latent variables – or scales
- Determine the dimensionality of these underlying latent variables

Factor Analysis

- Estimate underlying latent variables – or scales
- Determine the dimensionality of these underlying latent variables
- Recover measures of these underlying latent variables

Factor Loadings on the Unobserved Factors

- Consider a survey with i respondents who answer j survey questions

$$x_{ij} = \lambda_{j1}\xi_{i1} + \lambda_{j2}\xi_{i2} + \dots + \lambda_{jp}\xi_{ip} + \delta_{ij} \quad (2)$$

Factor Loadings on the Unobserved Factors

- Consider a survey with i respondents who answer j survey questions
- Factor analysis posits that x_{ij} is a combination of p unobserved factors, each written using the Greek letter ξ

$$x_{ij} = \lambda_{j1}\xi_{i1} + \lambda_{j2}\xi_{i2} + \dots + \lambda_{jp}\xi_{ip} + \delta_{ij} \quad (2)$$

Factor Loadings on the Unobserved Factors

- Consider a survey with i respondents who answer j survey questions
- Factor analysis posits that x_{ij} is a combination of p unobserved factors, each written using the Greek letter ξ

$$x_{ij} = \lambda_{j1}\xi_{i1} + \lambda_{j2}\xi_{i2} + \dots + \lambda_{jp}\xi_{ip} + \delta_{ij} \quad (2)$$

- λ are factor loadings

Factor Loadings on the Unobserved Factors

- Consider a survey with i respondents who answer j survey questions
- Factor analysis posits that x_{ij} is a combination of p unobserved factors, each written using the Greek letter ξ

$$x_{ij} = \lambda_{j1}\xi_{i1} + \lambda_{j2}\xi_{i2} + \dots + \lambda_{jp}\xi_{ip} + \delta_{ij} \quad (2)$$

- λ are factor loadings
- δ_{ij} is measurement error

Cronbach's Alpha: Measure of Scale Reliability

- Measure of internal consistency - how closely related a set of items are as a group

$$\alpha = \frac{N * \bar{c}}{\bar{v} + (N - 1) * \bar{c}} \quad (3)$$

Cronbach's Alpha: Measure of Scale Reliability

- Measure of internal consistency - how closely related a set of items are as a group
- is a function of the number of test item (N), the average inter-correlation among the items (\bar{c}), and the average variance (\bar{v})

$$\alpha = \frac{N * \bar{c}}{\bar{v} + (N - 1) * \bar{c}} \quad (3)$$

Factor Analysis of Trust in Political System Items

```
. factor $trust [aw=weight], pcf
(sum of wgt is 4.6914e+04)
(obs=45155)
```

Factor analysis/correlation

Method: principal-component factors
Rotation: (unrotated)

Number of obs = 45155
Retained factors = 2
Number of params = 13

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.24868	3.24066	0.6070	0.6070
Factor2	1.00803	0.28532	0.1440	0.7510
Factor3	0.72270	0.34281	0.1032	0.8542
Factor4	0.37989	0.11811	0.0543	0.9085
Factor5	0.26178	0.02687	0.0374	0.9459
Factor6	0.23491	0.09090	0.0336	0.9794
Factor7	0.14401	.	0.0206	1.0000

LR test: independent vs. saturated: $\chi^2(21) = 2.1e+05$ Prob> $\chi^2 = 0.0000$

Factor Loadings

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
trstprl	0.8182	-0.2303	0.2776
trstlgl	0.7828	-0.4055	0.2228
trstplc	0.7112	-0.4431	0.2979
trstplt	0.8516	-0.0036	0.2748
trstprt	0.8350	0.0480	0.3005
trstep	0.7323	0.5475	0.1639
trstun	0.7085	0.5406	0.2058

Factor Scores

```
. predict trust1 trust2
(regression scoring assumed)
```

Scoring coefficients (method = regression; based on varimax rotated factors)

Variable	Factor1	Factor2
trstprl	0.29475	-0.04882
trstlgl	0.40122	-0.18648
trstplc	0.41261	-0.22581
trstplt	0.15483	0.12735
trstprt	0.11863	0.16375
trstep	-0.22131	0.52508
trstun	-0.22114	0.51622