Week 9: Measurement

POL-GA 3200
Quantitative Field Methods
Prof. Cyrus Samii
NYU Politics

April 1, 2014

Some Basic Principles

Type	**Examples**	**Pros**	**Cons**
Administrative/ naturally observed	Vote records Expenditures	Unobtrusive Real-world	Limited availability Overdetermined
Induced behavior	Researcher audits Tests & games	Moderately obtrusive Incentive compatible Reveal mechanisms	Expensive Artificial Obtrusive
Self-rep. behaviors	Vote choice Organization membership	Private behavior Specific	Obtrusive Recall error/bias Social desirability
Self-rep. attitudes, perceptions, & opinions	Efficacy Legitimacy Preferences	Private beliefs Specific Reveal mechanisms	Obtrusive Social desirability Top-of-the-head

- You have many types of measures from which to choose.

Some Basic Principles

Type	**Examples**	**Pros**	**Cons**
Administrative/ naturally observed	Vote records Expenditures	Unobtrusive Real-world	Limited availability Overdetermined
Induced behavior	Researcher audits Tests & games	Moderately obtrusive Incentive compatible Reveal mechanisms	Expensive Artificial Obtrusive
Self-rep. behaviors	Vote choice Organization membership	Private behavior Specific	Obtrusive Recall error/bias Social desirability
Self-rep. attitudes, perceptions, & opinions	Efficacy Legitimacy Preferences	Private beliefs Specific Reveal mechanisms	Obtrusive Social desirability Top-of-the-head

- You have many types of measures from which to choose.
- Measurement theorists focus on "reliability" and "validity."

  `http://www.socialresearchmethods.net/kb/measure.php`

Some Basic Principles

Type	Examples	Pros	Cons
Administrative/ naturally observed	Vote records Expenditures	Unobtrusive Real-world	Limited availability Overdetermined
Induced behavior	Researcher audits Tests & games	Moderately obtrusive Incentive compatible Reveal mechanisms	Expensive Artificial Obtrusive
Self-rep. behaviors	Vote choice Organization membership	Private behavior Specific	Obtrusive Recall error/bias Social desirability
Self-rep. attitudes, perceptions, & opinions	Efficacy Legitimacy Preferences	Private beliefs Specific Reveal mechanisms	Obtrusive Social desirability Top-of-the-head

- You have many types of measures from which to choose.
- Measurement theorists focus on “reliability” and “validity.”

  `http://www.socialresearchmethods.net/kb/measure.php`
- I consider construct validity, power, and comparability.

Some Basic Principles

We will look at examples taking basic measurement a step further:

- Observed behaviors, sometimes induced by investigators.
- Index-type measures that aggregate multiple items to get at hard-to-measure concepts.
- Specialized questioning techniques.

Designing Behavioral Measures

- Best way to learn about “behavioral measurement” is to become familiar with examples.
- Some greatest hits...

- Audits:
    - Bertrand et al. (2007) surprise driving tests for corruption.
    - Olken (2007; 2009) road quality audits by engineers for corruption.
    - Olken & Barron (2009) riding with truckers to observe bribes and corruption.

- Audits:
  - Bertrand et al. (2007) surprise driving tests for corruption.
  - Olken (2007; 2009) road quality audits by engineers for corruption.
  - Olken & Barron (2009) riding with truckers to observe bribes and corruption.
- Correspondence:
  - Bertrand & Mullainathan (2004) CVs to measure discrimination.

- Audits:
  - Bertrand et al. (2007) surprise driving tests for corruption.
  - Olken (2007; 2009) road quality audits by engineers for corruption.
  - Olken & Barron (2009) riding with truckers to observe bribes and corruption.
- Correspondence:
  - Bertrand & Mullainathan (2004) CVs to measure discrimination.
- Economic transactions:
  - Dovidio et al (2010) attempts to buy $10 gift certificates with a check for discrimination.
  - Humphreys et al. (2012) and Beath et al. (2012) a wave of aid distribution for elite capture.

- Audits:
  - Bertrand et al. (2007) surprise driving tests for corruption.
  - Olken (2007; 2009) road quality audits by engineers for corruption.
  - Olken & Barron (2009) riding with truckers to observe bribes and corruption.
- Correspondence:
  - Bertrand & Mullainathan (2004) CVs to measure discrimination.
- Economic transactions:
  - Dovidio et al (2010) attempts to buy $10 gift certificates with a check for discrimination.
  - Humphreys et al. (2012) and Beath et al. (2012) a wave of aid distribution for elite capture.
- Games:
  - Glaeser et al. (2000) for trust.
  - Habyarimana et al. (2007) for ethnic preferences and institutions
  - Henrich et al. (2007) for trust and reciprocity.
  - Gilligan et al. (2013) for social capital.

- Audits:
    - Bertrand et al. (2007) surprise driving tests for corruption.
    - Olken (2007; 2009) road quality audits by engineers for corruption.
    - Olken & Barron (2009) riding with truckers to observe bribes and corruption.
- Correspondence:
    - Bertrand & Mullainathan (2004) CVs to measure discrimination.
- Economic transactions:
    - Dovidio et al (2010) attempts to buy $10 gift certificates with a check for discrimination.
    - Humphreys et al. (2012) and Beath et al. (2012) a wave of aid distribution for elite capture.
- Games:
    - Glaeser et al. (2000) for trust.
    - Habyarimana et al. (2007) for ethnic preferences and institutions
    - Henrich et al. (2007) for trust and reciprocity.
    - Gilligan et al. (2013) for social capital.
- Physical Markers:
    - Nisbett and Cohen (1996) cortisol response to an insult and physical contact aversion to measure aggressiveness.

Designing Behavioral Measures

- No recipe.
- Require creativity and adaptation to context and research question.
- Behavioral econ and social psych are rich in such measures.
- A concern is that context-specific measures are incomparable across studies.

Designing Indices

Indices measure attributes not easily observed or hard to capture by any single item:

Designing Indices

Indices measure attributes not easily observed or hard to capture by any single item:

- "Living standards" for judging extent and severity of poverty.

Designing Indices

Indices measure attributes not easily observed or hard to capture by any single item:

- "Living standards" for judging extent and severity of poverty.
- Psychological health and personality measures (e.g., self-esteem, "Big Five," PTSD).

Designing Indices

Indices measure attributes not easily observed or hard to capture by any single item:

- “Living standards” for judging extent and severity of poverty.
- Psychological health and personality measures (e.g., self-esteem, “Big Five,” PTSD).
- Customized/ad hoc approaches.

Designing Indices

Example of living standards for poverty
(Grosh & Glewwe, 2000; Deaton, 1997, Ch. 3):

Designing Indices

Example of living standards for poverty
(Grosh & Glewwe, 2000; Deaton, 1997, Ch. 3):

- Dominant approach: “money metric utility” approximated by consumption expenditure.
  - Consumption preferred over income as a measure due to consumption smoothing and variable income.
  - Smoothing within household means living standards best measured at household level (can divide by adult “consumption units”).
  - Convert consumption to $ metric w/ price index: $C_i = \sum_k p_k c_k$.
  - These arguments don’t apply if income *per se* interests you!

Designing Indices

Example of living standards for poverty
(Grosh & Glewwe, 2000; Deaton, 1997, Ch. 3):

- Alternatives: income or caloric intake.
- Literature is thick on alternative measures of living standards (cf. Ravallion, 2011, for a recent discussion).
- Deaton discusses aggregating LS to get at poverty, inequality, and other social welfare measures.
- Example: `LSMS (Grosh & Glewwe, 2000, pp.  31-46)`

Designing Indices

Example of Self-Esteem and Big Five:

- Commonly-used instrument for self-esteem is the “Rosenberg self-esteem scale” (lots of refs. on the web).
- Set of Likert items that are summed into a Likert scale: `http://www.wwnorton.com/college/psych/psychsci/media/rosenberg.htm`

Designing Indices

Example of Self-Esteem and Big Five:

- Commonly-used instrument for self-esteem is the “Rosenberg self-esteem scale” (lots of refs. on the web).
- Set of Likert items that are summed into a Likert scale:

  `http://www.wwnorton.com/college/psych/psychsci/media/rosenberg.htm`
- Big Five is similar, consisting of five Likert scales measuring openness, conscientiousness, extraversion, agreeableness, and neuroticism.

  `http://www.ocf.berkeley.edu/~johnlab/measures.htm`

  Poli sci application:

  `http://isps.research.yale.edu/publication/ISPS11-001/`

Designing Indices

- Common measure of internal consistency for sum scale, X:

$$\text{Cronbach's } \alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{k=1}^{K} \sigma_{Y_k}^2}{\sigma_X^2}\right),$$

  where $X = \sum_k Y_k$.
- Difference between σ_X^2 and $\sum_{k=1}^{K} \sigma_{Y_k}^2$ is covariance terms in σ_X^2.
- $\sigma_X^2 = \sum_{k=1}^{K} \sigma_{Y_k}^2$ implies no covariance, $\alpha = 0$.
- Convention is to rate $\alpha \geq 0.7$ acceptable.

Designing Indices

- Mokken score is an alternative (Van der Ark, 2007):

$$H = \frac{\sum_k \mathrm{Cov}\left(Y_k, X_{-k}\right)}{\sum_k c_k},$$

  where X_{-k} is $X - Y_k$, and c_k is a normalization coefficient measuring maximum attainable covariance.
- Convention is to rate $H \geq .4$ acceptable.
- Mokken analysis can also be used to check monotonicity in relationships between variables.

Designing Indices

Other kinds of indexing rules are also possible. Example of PTSD:

- A common screening tool used internationally is the WHO Composite International Diagnostic Interview. http://www.hcp.med.harvard.edu/wmhcidi/
- Scored as 0 or 1 depending based on Psych Diagnostic & Statistical Manual criteria:

  http://www.dsm5.org/Pages/Default.aspx

http://www.neurosurvival.ca/ClinicalAssistant/scales/dsm_IV/Anxiety.html

Designing Indices

Customized/ad hoc approaches:

- Indices we have seen were either weighted sum scores, based on pre-determined weights (e.g., prices), or simple sum scores.
- We may wish to come up with a way to weight items based on “information content” or extract “latent factors.”

Designing Indices

- *Inverse covariance weighting* optimizes information content for index constructed from *items determined to be related a priori*. Equiv. to a single factor latent variable model:

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{Ki} \end{pmatrix} = \begin{pmatrix} z_i + \varepsilon_{1i} \\ \vdots \\ z_i + \varepsilon_{Ki} \end{pmatrix}$$

Designing Indices

- *Factors scores* or *principal component scores* isolate and extract shared variation in *different* latent dimensions. Equivalent to a multifactor linear latent variable model with orthogonal factors:

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{Ki} \end{pmatrix} = \begin{pmatrix} \beta_1 z_{1i} + \ldots + \beta_K z_{Ki} + \nu_{1i} \\ \vdots \\ \beta_1 z_{1i} + \ldots + \beta_K z_{Ki} + \nu_{Ki} \end{pmatrix}$$

  where $\mathbf{z}_k' \mathbf{z}_l = 0$ for all $k \neq l$.

  `Look at R example...`

Designing Indices

- *IRT models* allow for similar index construction/factor extraction with binary, ordered, or categorical data, accounting for non-linearities between the linear combination of factors and observed data.

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{Ki} \end{pmatrix} = \begin{pmatrix} f(\beta_1 z_{1i} + \ldots + \beta_K z_{Ki}) \\ \vdots \\ f(\beta_1 z_{1i} + \ldots + \beta_K z_{Ki}) \end{pmatrix}$$

Designing Indices

Table 3: Main results

	(1)	(2)	(3)	(4)	(5)	(6)
	Lottery risk	Dictator	Cooperate	Trust sent	Trust return	Soc. Index
Violence	-0.11	2.04	0.16^{**}	1.68^{**}	0.07^{*}	0.57^{***}
	(0.24)	(1.35)	(0.07)	(0.63)	(0.03)	(0.13)
Observations	252	252	252	124	128	252
R^2	0.033	0.075	0.058	0.139	0.124	0.163
Baseline (no violence)	2.53	15.28	0.60	4.82	0.23	0.00

Standard errors in parentheses.

WLS with matched-pair block FE.

Robust standard errors clustered by ward. (p-values are for two-sided tests.)

Soc. Index is inverse covariance weighted average of outcomes 2-5.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Specialized Survey Techniques

Sensitive questions: Non-response or lying can be a problem. Tourangeau and Yan (2007) review evidence on ways to address this:

Specialized Survey Techniques

Sensitive questions: Non-response or lying can be a problem. Tourangeau and Yan (2007) review evidence on ways to address this:

- Survey administration:
    - Self-administered questionnaires with sealed ballot box, giving privacy.*
    - Interviewers who will be perceived as sympathetic.
    - “Bogus pipeline”: say you have a way to validate/detect lying.*

Specialized Survey Techniques

Sensitive questions: Non-response or lying can be a problem. Tourangeau and Yan (2007) review evidence on ways to address this:

- Survey administration:
  - Self-administered questionnaires with sealed ballot box, giving privacy.*
  - Interviewers who will be perceived as sympathetic.
  - “Bogus pipeline”: say you have a way to validate/detect lying.*
- Questioning techniques:
  - Using *forgiving* and *familiar* wording.
  - Prime people to have an honesty motive.*
  - Randomized response.*
  - Item count/list experiment.
  - Endorsement experiment (cf. work by Lyall et al.).
  - “Three card method” (illegal alien example).
  - Nominative method (“how many friends do you have who...”, cf. Salganik et al.).

Specialized Survey Techniques

Sensitive questions: Non-response or lying can be a problem. Tourangeau and Yan (2007) review evidence on ways to address this:

- Survey administration:
  - Self-administered questionnaires with sealed ballot box, giving privacy.*
  - Interviewers who will be perceived as sympathetic.
  - "Bogus pipeline": say you have a way to validate/detect lying.*
- Questioning techniques:
  - Using *forgiving* and *familiar* wording.
  - Prime people to have an honesty motive.*
  - Randomized response.*
  - Item count/list experiment.
  - Endorsement experiment (cf. work by Lyall et al.).
  - "Three card method" (illegal alien example).
  - Nominative method ("how many friends do you have who...", cf. Salganik et al.).
- *T&Y find consistent evidence in favor of these. Others either untested or inconsistent, meaning more research needed.

Specialized Survey Techniques

Other techniques that are out there:

- Anchoring vignettes (King et al., 2004): used to enhance interpersonal and inter-group comparability of expressed attitudes.
- Visual aids: Show cards are very common; More advanced techniques—e.g., using a pile of beans for respondents to elicit subjective probabilities (Delavande et al., 2010).

Remarks

- Many studies do very well in terms of identification and treatment construction, but then fail with measurement.

Remarks

- Many studies do very well in terms of identification and treatment construction, but then fail with measurement.
- Criteria for judging a measure are:
  - construct validity with respect to outcome of theoretical interest;
  - power for detecting effects;
  - comparability across studies ("mature" sciences work hard to establish standardized measures so that we can compare findings across studies, even do meta-analysis).

Remarks

- Many studies do very well in terms of identification and treatment construction, but then fail with measurement.
- Criteria for judging a measure are:
  - construct validity with respect to outcome of theoretical interest;
  - power for detecting effects;
  - comparability across studies ("mature" sciences work hard to establish standardized measures so that we can compare findings across studies, even do meta-analysis).
- Another criterion is "verifiability" (Blattman et al., 2014).

Remarks

- Many studies do very well in terms of identification and treatment construction, but then fail with measurement.
- Criteria for judging a measure are:
    - construct validity with respect to outcome of theoretical interest;
    - power for detecting effects;
    - comparability across studies ("mature" sciences work hard to establish standardized measures so that we can compare findings across studies, even do meta-analysis).
- Another criterion is "verifiability" (Blattman et al., 2014).
- Key trade-off is between validity of a measure in a given context and ability to be compare across contexts.