

# Measurement in the Social Sciences 1: Introduction and Summated Rating Model

Dave Armstrong

University of Oxford  
Department of Politics and International Relations  
Center for Research Methods in the Social Sciences

t: 01865 285956  
e: david.armstrong@politics.ox.ac.uk  
w: <http://www.quantoid.net/Oxford.php>

Office: 1101 Manor Road Building (Politics Dept)

April 30, 2009

1/37

## What You Will Learn Today

- What do we mean by “measurement”?
- What sort of theoretical models can we apply to existing data to generate “better” measures?
- One specific type of operational model - the Summated Rating Model.
  - What is Reliability and how do we estimate it?
  - How do we assess the fit of our data to the SRM?
  - How do we construct the measures?

2/37

## What Exactly Are We Doing?

- Statistically: Trying to use existing measures of the same underlying concept to generate “better” measures.
- Theoretically: Trying to obtain more precise, often more nuanced, and generally less error-laden measures of concepts of interest.

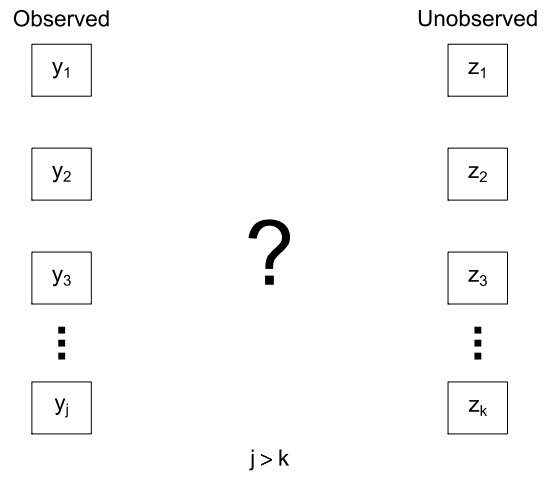
3/37

## Why do this?

- “Better” measure of some underlying concept, where “better” means:
  - Less polluted by measurement error.
  - Higher level of measurement.
  - Better able to distinguish between observations.
- Dimension Reduction.
- Reduce the effects of Multi-collinearity.

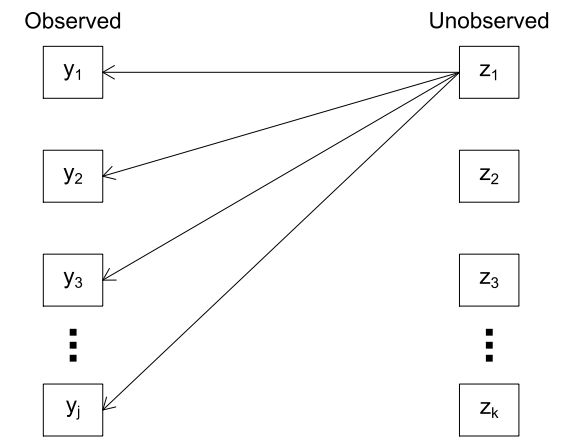
4/37

## Motivation



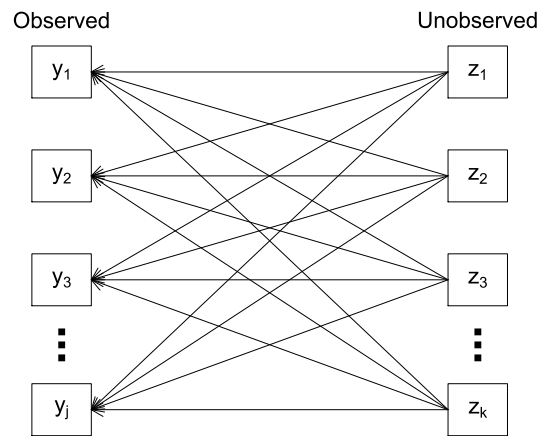
5/37

## Simple



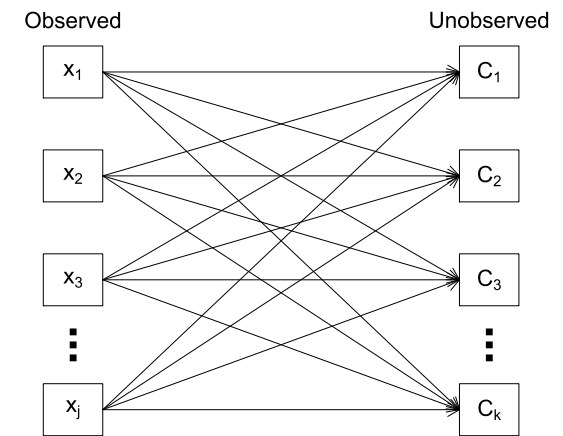
6/37

## Complex



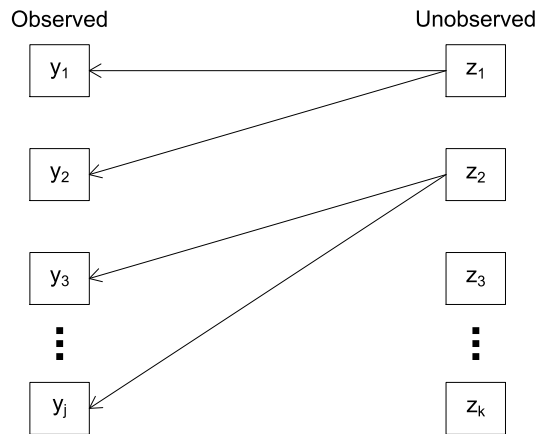
7/37

## Complex (2)



8/37

## Something in Between



9/37

## What we need

**Theoretical Model** A theoretical understanding and/or hypothesis and/or assumption about how observed and unobserved variables are related.

**Operational Model** A method of getting estimates of the parameters in the theoretical model.

10/37

## Glossary

**Measurement Error** Whatever makes an observation's value on a variable different from the "true" value.

**Parameter** Anything that requires estimation to obtain a numerical value.

**Latent Variable** (a.k.a. Unobserved Variable, Underlying Variable/Concept) Some variable that is unobserved and/or inherently unobservable.

**Vector** A series of numbers in a particular order. (e.g., a vector of coefficients, a variable vector). Can be a row or column vector.

**Matrix** A concatenation of a set of vectors with the same length, Indexed by [r,c].

11/37

## Vectors and Matrices

Column Vector:  $c = \begin{bmatrix} .1 \\ .2 \\ .3 \\ .4 \end{bmatrix}$  where  $c[1]=0.1$ ,  $c[4]=0.4$

Row Vector:  $r = [ .1 \ .2 \ .3 \ .4 ]$  where  $r[1]=0.1$ ,  $r[3] = 0.3$

Matrix:  $m = \begin{bmatrix} .1 & .4 & .7 \\ .2 & .5 & .8 \\ .3 & .6 & .9 \end{bmatrix}$  where  $m[1,2] = 0.4$ ,  $m[3,3] = 0.9$

12/37

## A math(s) primer: Summation Notation

$$\sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$$

$$\sum_{i=1}^n kX_i = k \sum_{i=1}^n x_i = k(X_1 + X_2 + \cdots + X_n)$$

$$\sum_{i=1}^n (X_i + Y_i) = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$$

$$= (X_1 + X_2 + \cdots + X_n) + (Y_1 + Y_2 + \cdots + Y_n)$$

$$\sum_{i=1}^n k = nk$$

$$\sum_{i=1}^n (X_i Y_i) \neq \sum_{i=1}^n X_i \sum_{i=1}^n Y_i$$

13/37

## A math(s) primer: Expectation Operator

The expectation operator is (for our purposes) the same as the mean. Technically:

$$E(X) = \sum_{i=1}^n p_i x_i$$

However, take a variable  $X$  with  $n$  observations where each  $x_i$  has a different value, then  $p_i = \frac{1}{n}$ . If we substitute this back into the equation above:

$$E(X) = \sum_{i=1}^n \frac{1}{n} x_i$$

$$= \frac{1}{n} \sum_{i=1}^n x_i$$

$$= \bar{X}$$

14/37

## A math(s) primer: Rules of the Expectation Operator

$$E(X + Y) = E(X) + E(Y)$$

$$E(X + k) = E(X) + k$$

$$E(kX) = kE(X)$$

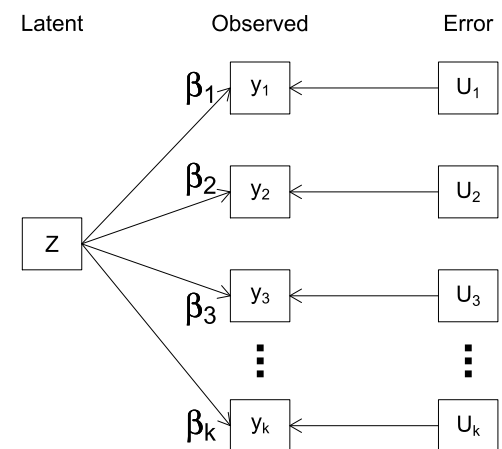
$$E(XY) \neq E(X)E(Y)$$

Variance of a Linear Combination:

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2abcov(X, Y)$$

15/37

## Simple model



16/37

## Simple model: Equation form

$$\begin{aligned}y_{i1} &= \beta_1 z_i + \varepsilon_{i1} \\y_{i2} &= \beta_2 z_i + \varepsilon_{i2} \\y_{i3} &= \beta_3 z_i + \varepsilon_{i3} \\&\vdots = \quad \quad \quad \vdots \\y_{ij} &= \beta_j z_i + \varepsilon_{ij}\end{aligned}$$

- How many data points do we have?
- How many parameters are there in this model?
- Note, from here on out, we are assuming that the observed variables have been standardized to have 0 mean and variance=1.

17/37

## Simplifying Assumptions

- $\beta_1 = \beta_2 = \beta_3 = \dots = \beta_j = 1$ , which gives us the following:

$$\begin{aligned}y_{i1} &= z_i + \varepsilon_{i1} \\y_{i2} &= z_i + \varepsilon_{i2} \\y_{i3} &= z_i + \varepsilon_{i3} \\&\quad \quad \quad \vdots = \quad \quad \quad \vdots \\y_{ij} &= z_i + \varepsilon_{ij}\end{aligned}$$

- There is some error that keeps the observed variable from being a perfect reflection of the “true” score on  $z$ .

18/37

## One Observation

- Now, let's look what happens to one observation  $i$ . For simplicity of notation here, let  $y_{i1}, y_{i2}, \dots, y_{i3} = y_1, y_2, \dots, y_3 = Y$ ,  $\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{ij} = \varepsilon_1, \varepsilon_2, \dots, \varepsilon_j = U$ , and  $z_i = z$ .

$$\begin{aligned}Y &= z + U \\E(Y) &= E(z + U) \\&= z + E(U)\end{aligned}$$

- Remember,  $E(U) = \bar{\varepsilon}$ . If we can assume that the errors cancel out, that is that the sum of the errors is 0 (making the mean of the errors also 0), then we have the following result:

$$E(Y) = z$$

19/37

## Is this model appropriate?

- Have to assume that the “signal” is the same across all of the variables in this model.
- Have to assume that there is no systematic measurement error. If every observation is off in the same direction, the estimate of the “true” dimension will also be biased, even in the limit.
- Also have to make a couple of other assumptions, but we'll get into those later.

20/37

## Reliability

- Let's look at one observed variable:  $Y_j = z + U_j$ . One thing that we're often interested in is how can we account for the variance in a variable:

$$\begin{aligned}\text{var}(Y_j) &= \text{var}(z + U_j) \\ &= \text{var}(z) + \text{var}(U_j) + 2\text{cov}(z, U_j)\end{aligned}$$

- If we now assume that  $\text{cov}(z, U_j) = 0$ , that is, the errors are independent of  $z$  and we'll also assume  $\text{cov}(e_i e_j) = 0 \forall i \neq j$  ( $\forall$  just means "for every").

$$\text{var}(Y_j) = \text{var}(z) + \text{var}(U_j)$$

21 / 37

## Reliability II

- Now, let's do a bit of rearranging:

$$1 = \frac{\text{var}(z)}{\text{var}(Y_j)} + \frac{\text{var}(U_j)}{\text{var}(Y_j)}$$

- Anybody know what this looks like?

$$1 = \frac{SS_{Reg}}{SS_{Tot}} + \frac{SS_{Err}}{SS_{Tot}}$$

- The first piece of the first equation on this page, is then like an  $R^2$  between the unobserved "true" score and the observed variable  $Y_j$ . We call this "reliability" - the proportion of variance in the observed variable that is due to the true, but unobservable dimension.
- In general, reliability refers to the repeatability and consistency of the measurement instrument. Thus, a measure is reliable if, when repeated, it produces similar results.

22 / 37

## What does Reliability mean?

- The higher a variable's reliability, the better a measure it is of the estimated underlying dimension.
- If a variable has very low reliability, it is unlikely that it is measuring the same thing as the other variables (no matter how much it's name or operationalization would suggest that it is).

23 / 37

## Assumptions thus far (and one new one)

- $e_{ij} \sim iid$  (also no systematic measurement error).
- $z_i$  is the same for all  $y_{ij}$  (uni-dimensionality).

Now, I'm going to introduce one more assumption now: Monotone Homogeneity.

- Monotone Homogeneity means that the relationship between the observed variables and the true underlying dimension is monotonically increasing.
- By monotonically increasing, I mean that as the true dimension increases, the observed variable cannot decrease (though it could stay the same).
- This suggests that the input variables don't need to be interval-level, they only need to be ordinal.

24 / 37

## Theoretical Conclusions

- What does all of this mean?
  - It means, that if the assumptions of this model are reasonable, then you can get an estimate of the underlying dimension by adding up, or taking the mean, of the observed variables for each observation.
  - Remember, we found that:

$$E(Y) = z + \underbrace{E(U)}_0$$

- Now, we have to worry about estimating the reliability.

25/37

## Estimating Reliability

- We can never know a variable's true reliability since it depends, in part, on the variance of the true score.
- We can get an estimate of a scale's reliability. In Stata, we do this with Cronbach's  $\alpha$  [alpha].
- The formula for  $\alpha$  is as follows:

$$\alpha = \frac{k\bar{r}}{1 + \bar{r}(k - 1)}$$

where  $\bar{r}$  is the average of the elements below the diagonal of the correlation matrix for the observed variables and  $k$  is the number of observed variables.

26/37

## Reliability "by hand"

$$r = \begin{bmatrix} 1.00 & 0.79 & 0.81 & 0.80 \\ 0.79 & 1.00 & 0.79 & 0.81 \\ 0.81 & 0.79 & 1.00 & 0.83 \\ 0.80 & 0.81 & 0.83 & 1.00 \end{bmatrix}$$

$$\bar{r} = \frac{1}{6}(0.79 + 0.81 + 0.80 + 0.79 + 0.81 + 0.83) = 0.805, k=4,$$

$$\begin{aligned} \alpha &= \frac{4(.805)}{1 + 0.805(3)} \\ &= \frac{3.22}{3.42} \\ &= 0.94 \end{aligned}$$

27/37

## Reliability the easy way

Open the dataset `srm.dta` in Stata. Then type:

```
alpha X1-X4, std
```

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

```
Average interitem correlation:      0.8044  
Number of items in the scale:      4  
Scale reliability coefficient:      0.9427
```

We know that  $0 \leq \text{Reliability} \leq 1$  and the closer to 1, the more reliable the scale is. This scale is quite reliable.

28/37

## More definitions.

- It seems that we have a reliable scale, now we have to assess the extent to which our assumptions hold.

Assume that we have observed variables  $y_{i1}, y_{i2}, \dots, y_{ik}$

- The “test score” is simply  $\sum_{j=1}^k y_{ij}$ .
- The “rest score” for variable  $y_{i1}$  is simply  $\sum_{j=2}^k y_{ij}$ . This is sum of the *rest* of the variables. The rest score for  $y_{i5}$  would be  $\sum_{j=1}^4 y_{ij} + \sum_{j=6}^k y_{ij}$ .

29/37

## Diagnostics

Now, tell Stata to do the following: `alpha X1-X4, std item.`

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

Item	Obs	Sign	item-test correlation	item-rest correlation	average inter-item correlation	alpha
X1	1000	+	0.9204	0.8568	0.8085	0.9268
X2	1000	+	0.9176	0.8520	0.8119	0.9283
X3	1000	+	0.9273	0.8688	0.7999	0.9230
X4	1000	+	0.9296	0.8728	0.7971	0.9218
Test scale					0.8044	0.9427

30/37

## Reading the Stata Output

- The “item-test correlation” is the correlation between the individual variable and the test score (the scale created by all of the variables).
  - This may not be the best measure because the test score contains the variable of interest, so those two things are related by definition.
- The “item-rest correlation” is the correlation between the individual variable and that variable’s rest score (the scale made from summing all of the rest of the variables).
- The “average inter-item correlation” is the average correlation between the rest of the items in the scale.

31/37

## Output continued

- The “alpha” column shows what Cronbach’s  $\alpha$  would be if we omitted that variable from the scale. Ideally, you want the overall  $\alpha$  to be bigger than it would be if you deleted any of the individual items.
- The “sign” column shows the direction of the relationship between the variable and the proposed scale. This could be negative if, for instance, a question was worded negatively and the rest were worded positively.
  - You can use the `reverse()` option to tell Stata which variables should be in the negative direction. These findings should not be surprising and if they are, you should definitely go back and take a look at your data.

32/37

## Evaluating Assumptions

- Uni-dimensionality: To evaluate this assumption, it is probably best to look at the correlation matrix. What you want to see is relatively similar (hopefully high-ish) correlations across the matrix. To the extent the blocks of high and low correlations exist, that is a problem.
- Monotone homogeneity can be evaluated by plotting each variable against its rest score. I've written a Stata program that will do this called "restplot". The program just takes a variable list and then plots each variable against the row-mean of the remaining variables.
  - See the extended example for use and output of this program.

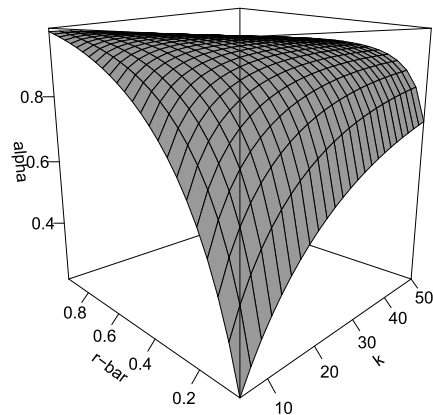
33/37

## Caveats and Cautions

- Cronbach's  $\alpha$  is not a good *test* of dimensionality. It is likely to give an underestimate of the dimensionality of your data. Testing dimensionality is a task better suited to factor analysis, which we will talk about later.
- As you include more variables, it is possible to see a relatively high  $\alpha$  value without very high inter-correlations.

34/37

$\alpha$  as a function of  $k$  and  $\bar{r}$



35/37

## SRM Conclusion

- At the end, you have a new variable (the sum of the observed variables for each observation) that is an estimate of the latent dimension.
- If the assumptions of this model hold, the estimate is "better" than any of the individual variables because the idiosyncracies and measurement error have canceled each other out.
- The resulting variable is (roughly) an interval level variable after starting with ordinal level input data.
- See Appendix 1 for extended examples.

36/37

## What You Learned Today

- What do we mean by “measurement”?
- What sort of theoretical models can we apply to existing data to generate “better” measures?
- One specific type of operational model - the Summated Rating Model.
  - What is Reliability and how do we estimate it?
  - How do we assess the fit of our data to the SRM?
  - How do we construct the measures?