Classical Test Theory, Reliability, and the Summated Rating Model

Measurement, Scaling, and Dimensional Analysis
2019 ICPSR Summer Program
Prof. Adam M. Enders

Objectives

- Account for, or represent, variability in a set of objects using a single dimension
  - That is, there is only one source of *important* or *substantively interesting* variability among a set of data
- Also want the resultant estimate of the dimension to be statistically reliable – more reliable than employing each individual item used to estimate the dimension
- If we achieve these two objectives, we will also have "better" measured the construct of theoretical interest

The Summated Rating Model

- Some people call them Likert (pronounced "lick-ert") scales, additive indexes (scales)
  - Note: Likert scale $\neq$ Likert response format!!!!
- Start with an $n \times k$ **rectangular** matrix of **dominance** data
  - Each variable, v_k is an imperfect measure of some characteristic
  - Each variable v_k is measured on the same scale
- General Procedure:
  - "Collapse" across the columns to end up with a single column vector of scores (calculate the mean within each row, or could calculate the sum)
  - If you have k items with m categories, the scale will have $k(m-1)+1$ distinct scores
  - Thats it!

Substantive Examples

- Racial resentment
- Political knowledge
- Political participation
- Values
- Issue attitudes
- Supreme Court legitimacy
- Partisanship (the Huddy and Greene social identity measures)
- Government spending preferences
- Racial stereotypes

The Summated Rating Model, cont'd

- It is equally possible to collapse across rows to get summary scores for the column objects
  - Might want to do this in order to learn something about the column objects
- Want to use this when you want to measure the variability in one set of objects (columns or rows) and not another
- We lose the information about the variability of the individual items when we use this model
- Most appropriate to use this when we want to measure a dimension we are pretty sure exists – *they are not good for testing dimensionality*
  - These models are prone to false positives – supporting a dimension when it doesn't exist
- A successful application of the summated rating scale **generates an interval-level estimate of the underlying dimension** from ordinal-level variables

The Summated Rating Model, cont'd

- Classical Test Theory: a psychometric theory of item response where responses can be broken down into "true scores" and (random) error
  - Also concerned with improving reliability of measurements, as we'll see shortly
- Some terminology:
  - T: "true" underlying dimension
  - $V_j, j = 1, 2..., k$: item
  - X: the scale formed from the V_j's
- Each item in an SRS is associated with a trace line/item characteristic curve/item response function
  - A trace line is a graph of $E(V_j)$ for each position along T
  - There will be k different trace lines (that is, one of these graphs for each item)
  - We can't empirically construct these graphs, though – we don't know T!

Model Assumptions

- **Single assumption of this model**: trace line for each item is monotonic with respect to the underlying dimension T (called the monotone homogeneity assumption)
  - These monotonic curves are the measurement functions of each of the individual items (since they are monotonic, its equivalent to say that we are working with ordinal items)
  - $X_i = \sum_{j=1}^{k} V_{ij}$: this is equivalent to summing across the trace lines
  - Taking $E(\sum_{j=1}^{k} V_{ij})$ will cancel out the idiosyncrasies in the trace lines, *leaving us with a linear trace line, or, an interval-level estimate*

Item Analysis

- Most people do not check the single assumption listed above: need to perform an **item analysis**
- Item analysis consists of checking the monotonicity of the trace lines associated with each item by plotting each variable involved in the scale against the true dimension
  - Again, this is difficult because we don't have T
- So, we examine bivariate relationships between the V_j's and X – most people correlate the item with the scale $r(V_j, X_j)$

Item Analysis, cont'd

- Problem: the correlation is biased upward if we correlate the a given item and a scale including the item
  - Need to correlate the item with a scale that excludes that item $r(V_j, X_{-j})$, called the **rest score**
  - *All correlations should be positive*, but the correlation need not be large or statistically significant (since each of the V_j's are measured with error, which attenuates correlation coefficients)
  - Furthermore, correlation coefficients in this instance are a measure of the *linear* relationship between the item and the true dimension; BUT, we only want to assume monotonicity, not linearity!
- Solution: fit a smoother to a scatter plot of V_j against the scale X_{-j}

Why use a scale in the first place?

- Why not just find the item with the "straightest" trace line?
- Consider the following: $V_j = T + e_j$, the response of an individual to item V_j is comprised of the their true ideal point T and measurement error, e_j
- We are going to assume that over repeated trials $E(e_j) = 0$
- Multiple items help us reduce the net effect of error

Reducing error

(Note, I have dropped the i subscript below for simplicity):

$$V_j = T + e_j$$

$$\sum_{j=1}^{k}(V_j) = \sum^{k}(T + e_j)$$

$$\frac{\sum_{j=1}^{k}(V_j)}{k} = \frac{\sum^{k}(T + e_j)}{k}$$

$$X = \frac{kT}{k} + \frac{\sum^{k} e_j}{k}$$

$$X = T + \bar{e}$$

Reducing error

- The scale score is a composite of the true score and the sample mean of the error (which should be close to 0)
- The Central Limit Theorem also tells us that the variance of the error, σ_e^2, is equal to $\frac{\sigma_e^2}{k}$ when we have k items
- So, error variance reduces as the number of items, k, increases (or, when we increase "scale length")
- This is why we use multiple-item scales!
  - To be clear: multiple item scales are always more reliable than single items
  - One of the best practices you can adopt when it comes to quantitative analysis is always employing multiple item measures (i.e., scales) of key variables of interest
  - If there aren't any in your field, develop one!

Reducing error

Let's assume: $E(e_i) = 0$, $COV(e_i, T_i) = 0$, and $COV(e_i, e_j) = 0$

$$X_i = \sum_{j=1}^{k} V_{ij} = T_i + e_i$$

$$\begin{aligned} X_i &= T_i + e_i \\ E(X_i) &= E(T_i + e_i) \\ E(X_i) &= T_i \end{aligned}$$

$$\begin{aligned} VAR(X_i) &= VAR(T_i + E_i) \\ &= VAR(T_i) + VAR(E_i) + 2COV(T_i, E_i) \\ 1 &= \frac{VAR(T_i)}{VAR(X_i)} + \frac{VAR(E_i)}{VAR(X_i)} \end{aligned}$$

Reliability

- $\frac{VAR(T)}{VAR(X)}$ = the **reliability** of X
  - It is the proportion of variability in the scale scores not due to error (or, alternatively, variance that can be attributed to the true scale)
- Reliability can be thought of as the squared correlation ($R^2_{X,T}$) between the scale scores and the "true," underlying scale
- The "tightness" or "compactness" of the data points around the best fitting trace line of the scale scores to the true scores is greater when there is a larger number of items

Parallel Measures

- So how can we calculate reliability if we don't have T (in fact, if we had T we wouldn't need to construct the scale!)?
- Need **parallel measures**
- Take X, X^* and X^{**} to be separate empirical scales, all of which measure T
- Following previous material:
  - $X_i = T_i + E_i$
  - $X_i^* = T_i + E_i^*$
  - $X_i^{**} = T_i + E_i^{**}$

Parallel Measures, cont'd

- X, X^* and X^{**} are parallel measures if the following conditions hold:
  1. $E(X) = E(X^*) = E(X^{**}) = E(T)$
  2. $\sigma^2_X = \sigma^2_{X^*} = \sigma^2_{X^{**}}$
  3. $\sigma_{X,X^*} = \sigma_{X,X^{**}} = \sigma_{X^*,X^{**}}$
  4. $\sigma_{X,Y} = \sigma_{X^*,Y} = \sigma_{X^{**},Y}$, where Y is some other variable, not a measure of T
- Part of number 1 and numbers 2-4 are empirically testable

Parallel Measures, cont'd

Now, consider correlation between two parallel measures:

$$
\begin{aligned}
\rho_{X,X^*} &= \frac{\sigma_{X,X^*}}{\sigma_X \sigma_{X^*}} \\
&= \frac{\sigma_{(T+E)(T+E^*)}}{\sigma_X \sigma_{X^*}} \\
&= \frac{\sigma_{(T^2+TE+TE^*+EE^*)}}{\sigma_X \sigma_{X^*}} \\
&= \frac{\sigma_{T^2} + \sigma_{TE} + \sigma_{TE^*} + \sigma_{EE^*}}{\sigma_X \sigma_{X^*}} \\
&= \frac{\sigma_T^2}{\sigma_X \sigma_{X^*}} \\
&= \frac{\sigma_T^2}{\sigma_X \sigma_X} \\
\rho_{X,X^*} &= \frac{\sigma_T^2}{\sigma_X^2} \longrightarrow \text{reliability}
\end{aligned}
$$

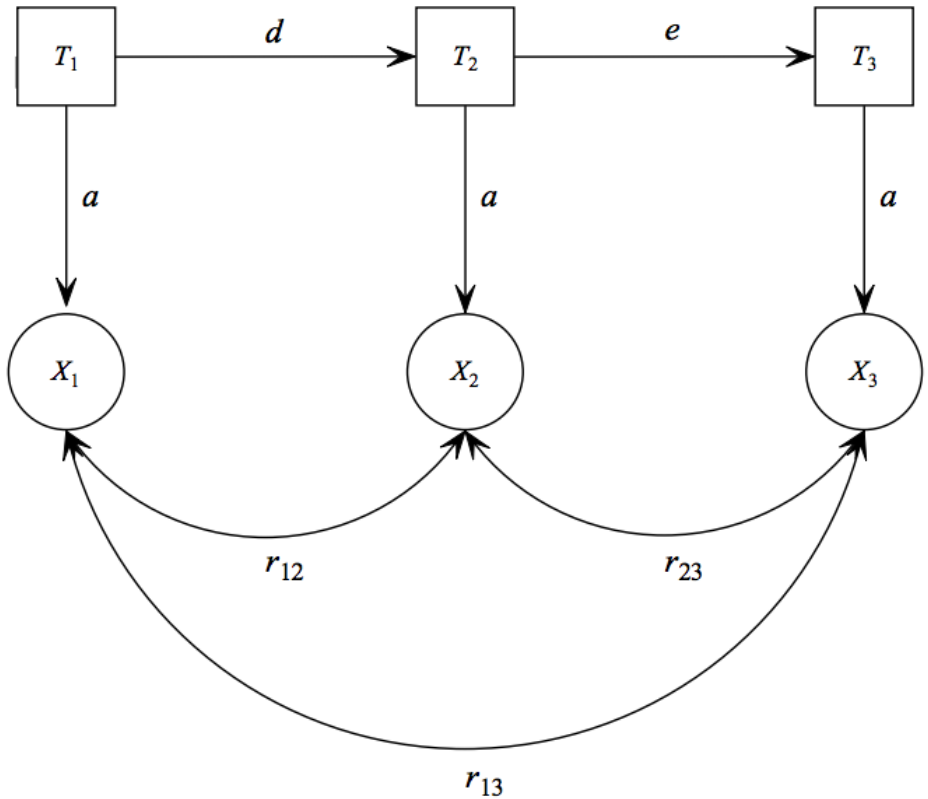
Where do We Get Parallel Measures?

- One source of a parallel measure is using a scale constructed at some other point in time (repeated measures over time)
- Some obvious problems:
  - Need panel data
  - The true scale scores could have changed over time, which would affect your inferences
  - Really, need at least 3 measures to be sure change isn't due to real change
- Easy method of calculating reliability with panel data produced by Heiser (1969) and further developed by Wiley and Wiley (1970)

Estimating Reliability from Panel Data

- X is a summated rating scale that is assumed to measure some underlying "true" dimension, T
- Assume that we have three-wave panel data that provide us with three successive values of X for the same observations
- Denote the scale scores and true scores at time point j as X_j and T_j , respectively, with $j = 1, 2, 3$

Estimating Reliability from Panel Data, cont'd



a, d, and e are coefficients to be estimated while r_{12}, r_{23}, and r_{13} are empirical correlations between the scale scores across the three waves of the panel

Estimating Reliability from Panel Data, cont'd

- The rules of path analysis can be used to express the correlations as functions of the coefficients, as follows:
  - $r_{12} = a^2 d$
  - $r_{23} = a^2 e$
  - $r_{13} = a^2 de$
- It follows, then, that:
  - $a^2 = \frac{r_{12} r_{23}}{r_{13}}$
  - $d = \frac{r_{13}}{r_{23}}$
  - $e = \frac{r_{13}}{r_{12}}$

An Example: Party Identification

- Though never referred to this by the authors of *The American Voter* (1960), partisanship has subsequently been colloquially referred to as the "unmoved mover"
- This implies stability, which itself implies relatively high reliability

	(1)	(2)	(3)
(1) Party ID_{92}	1.000		
(2) Party ID_{94}	0.800	1.000	
(3) Party ID_{96}	0.772	0.857	1.000

$$a^2 = \frac{0.800 \times 0.857}{0.772} = 0.888 \qquad d = \frac{0.772}{0.857} = 0.901 \qquad e = \frac{0.772}{0.800} = 0.965$$

Where do We Get Parallel Measures?, cont'd

- Equivalence strategy: two simultaneous measures of T
  - Split-half measure: divide items in half, construct two scales, and correlate them
    - Problem: if we divide the number of items in half, we will have an inherently biased (downward) measure of scale reliability
    - Spearman-Brown Prophecy formula will correct for reduced items: $r_{x^*,x^{**}} = \frac{2r}{1+r}$
    - Problem: there are lots of way we can divide the items in half
    - Could take the mean split-half correlation to correct this, but this would take a long time because you would have to construct all the scales
  - **Cronbach's α**: mean split-half correlation corrected for scale length

Cronbach's Alpha

- Typical formula:

$$\alpha = \frac{k\bar{r}}{1 + \bar{r}(k-1)}, \text{ where } \bar{r} = \text{mean correlation}$$

- This is a lower-bound estimate of the true reliability of the scale (not always a bad thing!)
- It underestimates because:
  1. We typically do not have perfectly parallel measures
  2. Our trace lines only require monotonicity
- Rest scores tell us how α changes when a given item is removed from the analysis – if α increases, that item probably doesn't belong in the model

Perils of Cronbach's Alpha

Alpha should **NOT** be used a measure of internal consistency or homogeneity

	V_1	V_2	V_3	V_4	V_5	V_6
V_1:	—					
V_2:	0.90	—				
V_3:	0.90	0.90	—			
V_4:	0.00	0.00	0.00	—		
V_5:	0.00	0.00	0.00	0.90	—	
V_6:	0.00	0.00	0.00	0.90	0.90	—

$$\alpha = \frac{k\bar{r}}{1 - \bar{r}(k-1)} = \frac{6 \times 0.36}{1 + (5 \times 0.36)} = \frac{2.16}{2.80} = 0.77$$

Perils of Cronbach's Alpha

Alpha also increases as k increases, regardless of average correlation, $\bar{r}$

Say we have a large scale with $k = 30$, but a small average correlation $\bar{r} = 0.075$

$$\alpha = \frac{k\bar{r}}{1 - \bar{r}(k-1)} = \frac{30 \times 0.075}{1 + (29 \times 0.075)} = \frac{2.25}{3.175} = 0.709$$

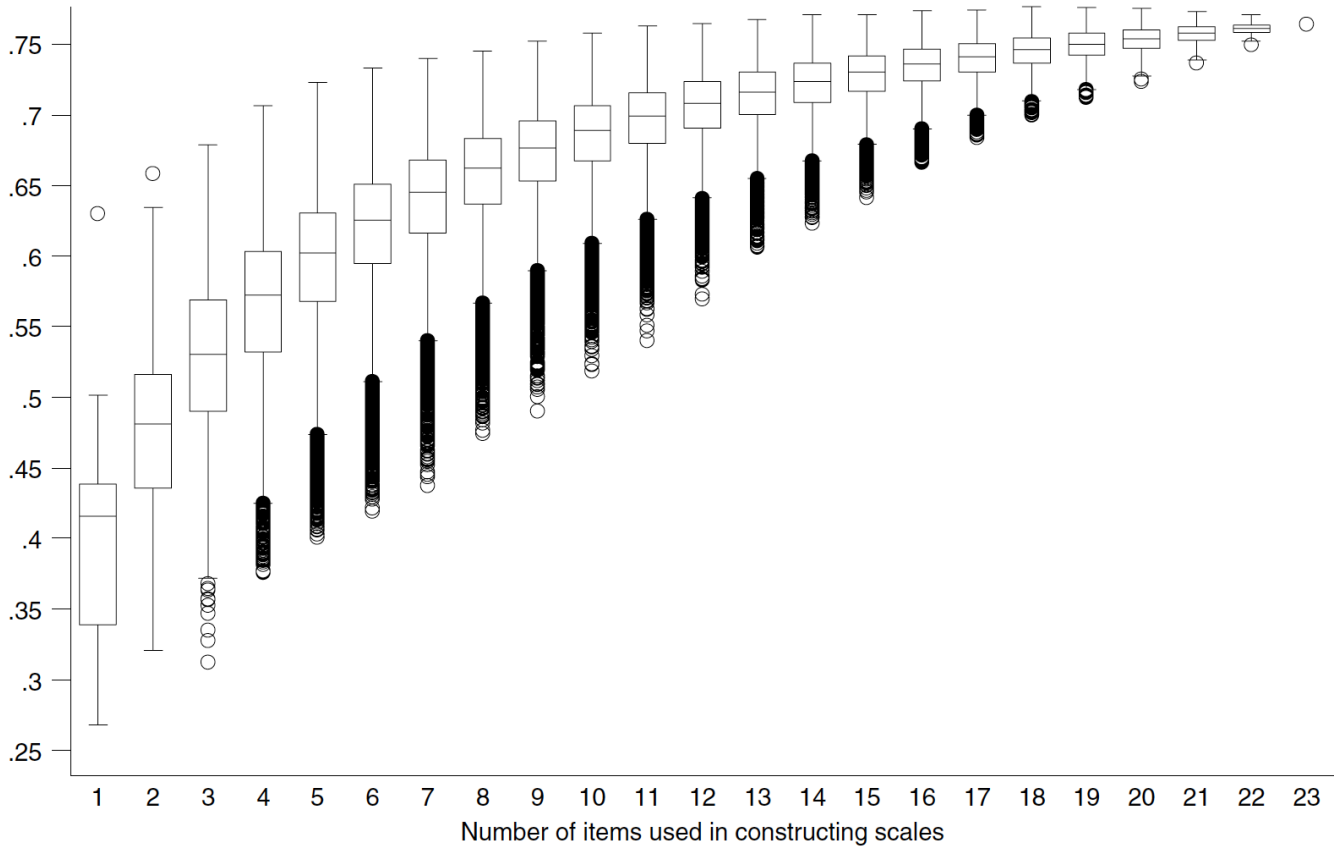
Used a measure of internal consistency, we would accept this as a "good" scale with items that "go together" – this is clearly not the case

Perils of Cronbach's Alpha, cont'd

- Alpha is smaller than the greatest lower bound
  - alpha $\leq$ GLB $\leq \rho_{X,X^*}$
  - Other types of reliability estimates that are perfectly good (e.g., Guttman's λ_2, which is larger than α but smaller than GLB)
- All that said: most people only know Cronbach's alpha and are going to ask for that during the review process #science
- Citations to Cronbach's 1951 *Psychometrika* paper outrank even Watson and Crick's 1953 *Nature* article describing their discovery of the double helix structure of DNA

Application: Ansolabahere et al. (2008)

FIGURE 1. Correlation Between 1990 and 1992 Economic Issue Scales Box-and-Whiskers Plot



Application: Ansolabahere et al. (2008)

TABLE 4. Correlations Between ANES Panel Waves, Issue Scales, and Individual Survey Items by Education and Political Information Level

	High-Educ. Respondents		Low-Educ. Respondents		High-Info. Respondents		Low-Info. Respondents	
Issue Area	Issue Scales	Indiv. Items	Issue Scales	Indiv. Items	Issue Scales	Indiv. Items	Issue Scales	Indiv. Items
1992, 1996								
Economic Issues	.81	.47	.71	.35	.78	.45	.68	.30
Moral Issues	.86	.58	.75	.42	.84	.54	.71	.38
Ideology		.84		.31		.73		.19[1]
Party ID		.77		.79		.81		.76
1990, 1992								
Economic Issues	.81	.50	.74	.36	.78	.46	.68	.31
Racial Issues	.86	.57	.73	.48	.82	.54	.68	.43
Moral Issues	.85	.57	.58	.35	.82	.53	.47	.31
Ideology		.79		.39		.67		.32
Party ID		.83		.76		.83		.57
1972, 1976								
Economic Issues	.73	.46	.66	.39	.71	.45	.60	.35
Racial Issues	.80	.44	.72	.37	.80	.47	.69	.32
Womens' Issues	.76	.47	.59	.36	.70	.45	.60	.36
Law & Order	.82	.61	.63	.45	.78	.58	.58	.36
Ideology		.66		.48		.63		.56
Party ID		.87		.71		.83		.61
1956, 1960								
Economic Issues	.62	.47	.58	.39				
Party ID		.92		.79				

Conclusions

- Although simply, the summated rating model is quite powerful
  - Reduces dimensionality of dataset, increases level of measurement
  - Results in more reliable measure of the construct we're interested in
  - Can be used to test substantive theories about nature of item response
- Only get the above things by thinking seriously about the model and assumptions
- Will continue this practice, introducing more/different assumptions in moving to the cumulative scaling model and IRT models