Measurement Reliability

Qian-Li Xue
Biostatistics Program
Harvard Catalyst | The Harvard Clinical & Translational Science Center

Short course, October 27, 2016

Objectives

- Classical Test Theory
- Definitions of Reliability
- Types of Reliability Coefficients
  - Test-Retest, Inter-Rater, Internal Consistency,
  - Correction for Attenuation
- Review Exercises

What is reliability

- Consistency of measurement
- The extent to which a measurement instrument can differentiate among subjects
- Reliability is relative

Facets of Reliability

- Mrs. Z scores 20 at visit 1 and 25 at visit 2. Could be:
- Random variation
  - **(Test-Retest)**
- Tech # 2 more lenient than Tech # 1
  - **(Inter-Rater Reliability)**
- Version # 2 easier than Version # 1
  - (Related to **Internal Consistency)**
- Mrs. Z's picture-naming actually improved

Classical Test Theory

- $X = T_x + e$
- The Observed Score = True Score + Error
- Assumptions:
  - $E(e) = 0$
  - $Cov(T_x, e) = 0$
  - $Cov(e_i, e_k) = 0$

- $Var(X) = Var(T_x + e) = Var(T_x) + 2Cov(T_x, e) + Var(e)$
- $Var(X) = Var(T_x) + Var(e)$

Reliability as Consistency of Measurement

- The relationship between parallel tests

- Ratio of True score variance to total score variance

$$\rho_{xx} = \frac{Var(T_x)}{Var(X)} = \frac{Var(X)-Var(e)}{Var(X)}$$

Parallel Tests

- Parallel: $T_{X_1} = T_{X_2} \quad Var(\varepsilon_1) = Var(\varepsilon_2)$
- Tau-Equivalent: $T_{X_1} = T_{X_2}$
- Essentially Tau-Equivalent: $T_{X_1} = T_{X_2} + c$
- Congeneric: $T_{X_1} = \beta T_{X_2} + c$

See Graham (2006) for details.

Correlation, *r*

Correlation (i.e. “Pearson” correlation) is a scaled version of covariance

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}$$

$-1 \leq r \leq 1$

$r = 1$	perfect positive correlation
$r = -1$	perfect negative correlation
$r = 0$	uncorrelated

Correlation between Parallel Tests

- $\rho_{X_1X_2}$ equal to reliability of each test

$$\rho_{X_1X_2} = \frac{\operatorname{cov}(T_{X_1} + \varepsilon_1, T_{X_2} + \varepsilon_2)}{\sqrt{\operatorname{var}(X_1)\operatorname{var}(X_2)}}$$

$$= \frac{\operatorname{cov}(T_{X_1}, T_{X_2}) + \operatorname{cov}(T_{X_1}, \varepsilon_2) + \operatorname{cov}(T_{X_2}, \varepsilon_1) + \operatorname{cov}(\varepsilon_1, \varepsilon_2)}{\sqrt{\operatorname{var}(X_1)\operatorname{var}(X_2)}}$$

$$\rho_{X_1X_2} = \frac{\operatorname{var}(T_X)}{\operatorname{var}(X)}$$

DIADS Example

- Depression in Alzheimers Disease Study.
- Placebo-controlled double-blind controlled trial of sertraline
- One outcome was the Boston Naming Test.
- Consists of 60 pictures to be named, two versions.

Measures for Reliability

	Continuous	Categorical
Test-retest	r or ICC	Kappa or ICC
Inter-rater	r or ICC	Kappa or ICC
Internal Consistency	Alpha or Split-half or ICC	KR-20 or ICC (dichotomous)

Kappa Coefficient
(Cohen, 1960)

- Test-Retest or Inter-rater reliability for categorical (typically dichotomous) data.
- Accounts for chance agreement

Observed		Rater 2		
		Present	Absent	Total
Rater 1	Present	20	15	35
	Absent	10	55	65
	Total	30	70	100

Kappa Coefficient

Observed		Rater 2		
		Present	Absent	Total
Rater 1	Present	20	15	35
	Absent	10	55	65
	Total	30	70	100

Expected		Rater 2		
		Present	Absent	Total
Rater 1	Present	10.5	24.5	35
	Absent	19.5	45.5	65
	Total	30	70	100

$$\text{kappa} = \frac{P_o - P_e}{1.0 - P_e}$$

P_o = observed proportion of agreements

P_e = expected proportion of agreements

$$\text{kappa} = \frac{[(20+55)/100]-[(10.5+45.5)/100]}{1-[(10.5+45.5)/100]} = 0.43$$

Kappa in STATA

Data Editor

Preserve | Restore | Sort | << | >>

patient[36] = 36

	patient	rater1	rater2
32	32	1	0
33	33	1	0
34	34	1	0
35	35	1	0
36	36	0	1
37	37	0	1
38	38	0	1
39	39	0	1
40	40	0	1
41	41	0	1
42	42	0	1
43	43	0	1

```
. kap rater1 rater2

                 Expected
Agreement       Agreement        Kappa    Std. Err.
-----------------------------------------------------
  75.00%          56.00%        0.4318     0.0994
```

Kappa Interpretation

- Interpretation:

Kappa Value	Interpretation
Below 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

(source: Landis, J. R. and Koch, G. G. 1977. *Biometrics* 33: 159-174)

- kappa could be high simply because marginal proportions are either very high or very low!!
- Best interpretation of kappa is to compare its values on other, similar scales

Weighted Kappa
(Cohen, 1968)

- For ordered polytomous data
- Requires assignment of a weighting matrix

Rater A	normal	benign	suspect	cancer
normal	1	.8	0	0
benign	.8	1	0	0
suspect	0	0	1	.8
cancer	0	0	.8	1

$$K_w = 1.0 - \frac{\sum w_{ij} \times P_{oij}}{P_{eij}}$$

- K_w=ICC with quadratic weights (Fleiss & Cohen, 1973)

Measures for Reliability

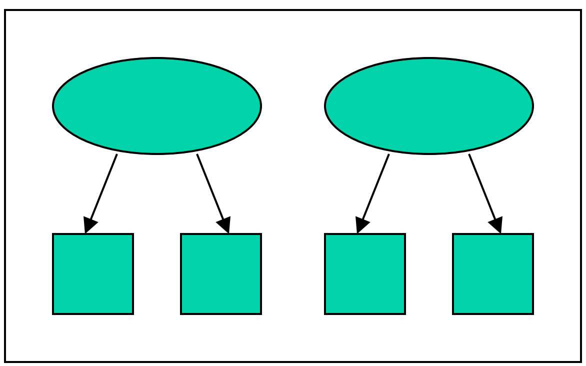
	Continuous	Categorical
Test-retest	r **or** ICC	Kappa **or** ICC
Inter-rater	r **or** ICC	Kappa **or** ICC
Internal Consistency	Alpha **or** Split-half **or** ICC	KR-20 **or** ICC **(dichotomous)**

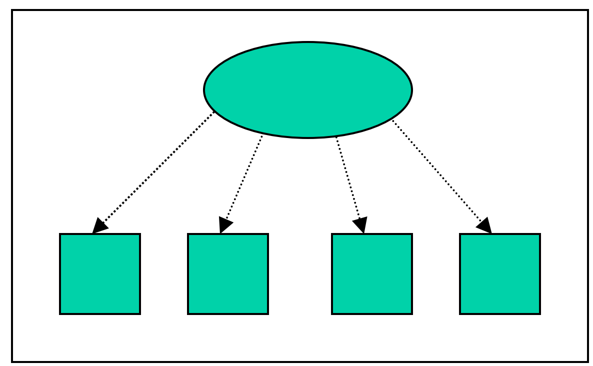
Internal Consistency

- Degree of homogeneity of items within a scale.
- Items should be correlated with each other and the total score.
- <u>Not</u> a measure of dimensionality; <u>assumes</u> unidimensionality.

Internal Consistency and Dimensionality

- Two (at least) explanations for lack of internal consistency among scale items:
  - More than one dimension
  - Bad items





Cronbach's Alpha

```
. corr Rating1 Rating2 Rating3 Rating4 Total, covariance
(obs=6)
```

	Rating1	Rating2	Rating3	Rating4	Total
Rating1	2.66667				
Rating2	2	2.7			
Rating3	1.93333	2.4	2.66667		
Rating4	3.06667	3	2.93333	6.26667	
Total	9.66667	10.1	9.93333	15.2667	44.9667

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} \sigma^2_{item_i}}{\sigma^2_{total}}\right] \quad \alpha = \frac{4}{3}\left[1 - \frac{2.67 + 2.7 + 2.67 + 6.27}{44.97}\right] = 0.91$$

Cronbach's Alpha

- Mathematically equivalent to ICC(3,*k*)

	single (,1)	mean (,*k*)
(1,) unique	0.17	0.44
(2,) random	0.29	0.62
(3,) fixed	0.71	$\boxed{0.91}$

- When inter-item correlations are equal across items, equal to the average of all split-half reliabilities.

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

See DeVellis pp 36-38

STATA Alpha Output

```
. alpha  Rating1 Rating2 Rating3 Rating4 ,item

Test scale = mean(unstandardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average inter-item covariance	alpha
Rating1	6	+	0.8828	0.8058	2.777778	0.8834
Rating2	6	+	0.9166	0.8593	2.644444	0.8665
Rating3	6	+	0.9071	0.8445	2.688889	0.8715
Rating4	6	+	0.9095	0.7902	2.111111	0.9179
Test scale					2.555556	0.9093

Kuder-Richardson 20

$$KR20 = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{\sigma^2_{total}}\right]$$

$p_i =$ Proportion responding positively to item *i*

$q_i = 1 - p_i$

- Cronbach's alpha for dichotomous items
- Use alpha command in STATA, will automatically give KR20 when items are dichotomous.

Correction for Attenuation

- You can calculate $r_{x,y}$
- You want to know r_{TxTy}

$$r_{T_x T_Y} = \frac{r_{x,y}}{r_{xx} r_{yy}}$$

Correction for Attenuation

observed correlation r(x,y) = 0.3					
	reliability of x r(x,x)				
r(y,y)	0.2	0.4	0.6	0.8	1
0.2			0.87	0.75	0.67
0.4		0.75	0.61	0.53	0.47
0.6	0.87	0.61	0.5	0.43	0.39
0.8	0.75	0.53	0.43	0.38	0.34
1	0.67	0.47	0.39	0.34	0.3

observed correlation r(x,y) = 0.5					
	reliability of x r(x,x)				
r(y,y)	0.2	0.4	0.6	0.8	1
0.2					
0.4				0.88	0.79
0.6			0.83	0.72	0.65
0.8		0.88	0.72	0.63	0.56
1		0.79	0.65	0.56	0.5

How to Improve Reliability

- Reduce error variance
  - Better observer training
  - Improve scale design
- Enhance true variance
  - Introduce new items better at capturing heterogeneity
  - Change item responses
- Increase number of items in a scale

Exercise #1

- You develop a new survey measure of depression based on a pilot sample that consists of 33% severely depressed, 33% mildly depressed, and 33% non-depressed. You are happy to discover that your measure has a high reliability of 0.90. Emboldened by your findings, you find funding and administer your survey to a nationally representative sample. However, you find that your reliability is now much lower. Why might have the reliability dropped?

Exercise #1 - Answer

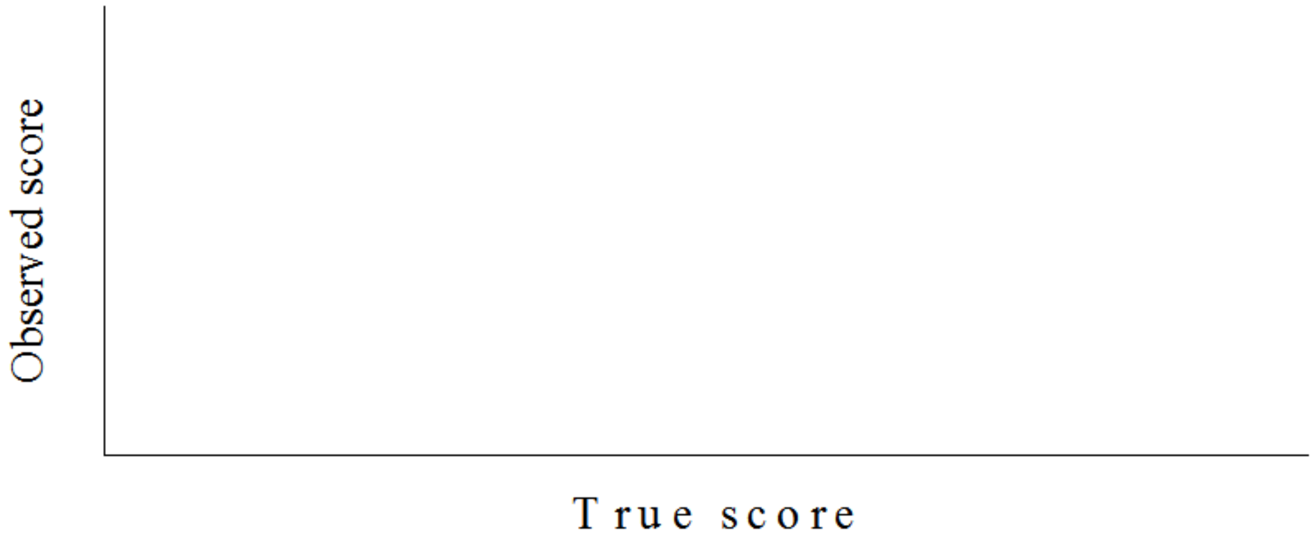
$$0.90 = \frac{BMS_{pilot} - EMS}{BMS_{pilot}} = \frac{10-1}{10}$$

$$ICC_{National} = \frac{BMS_{National} - EMS}{BMS_{National}} = \frac{4-1}{4} = 0.75$$

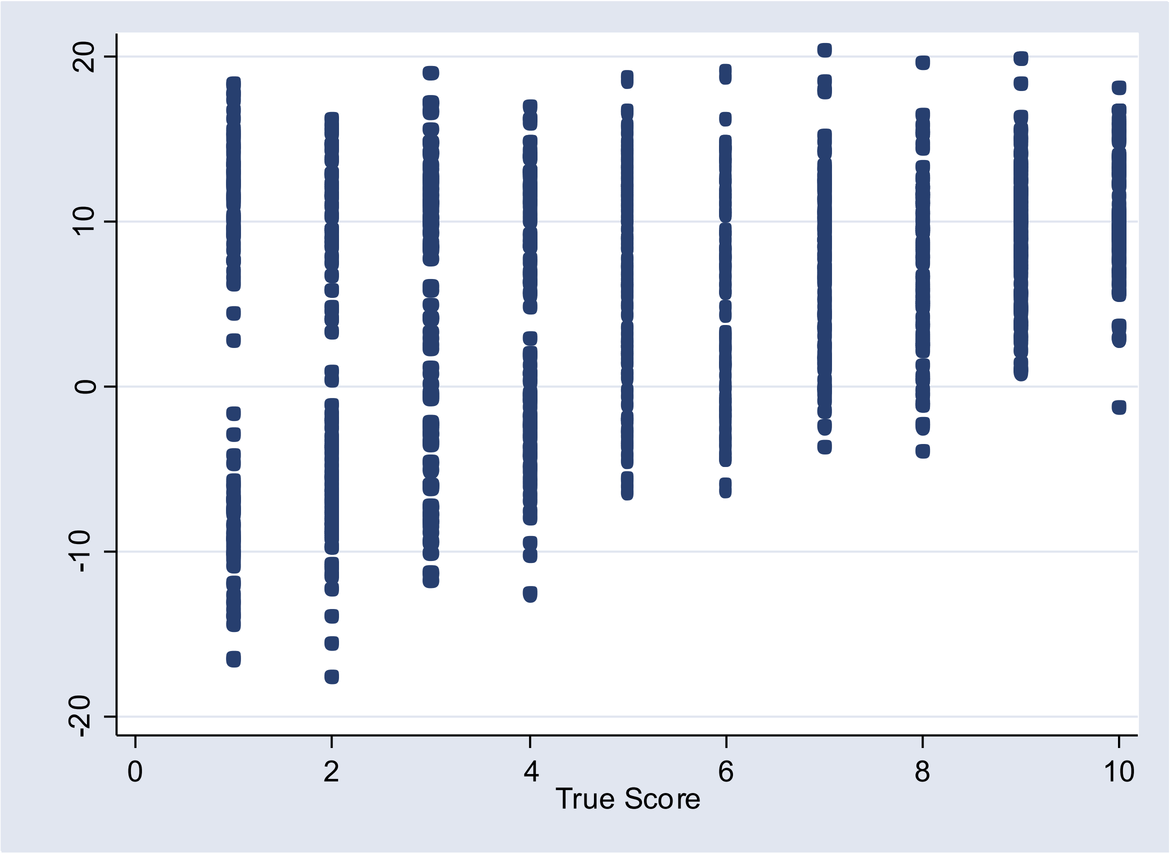
Suppose all of the national sample are severely depressed, then BMS (between-person variance) drops, as does ICC.

Exercise #2

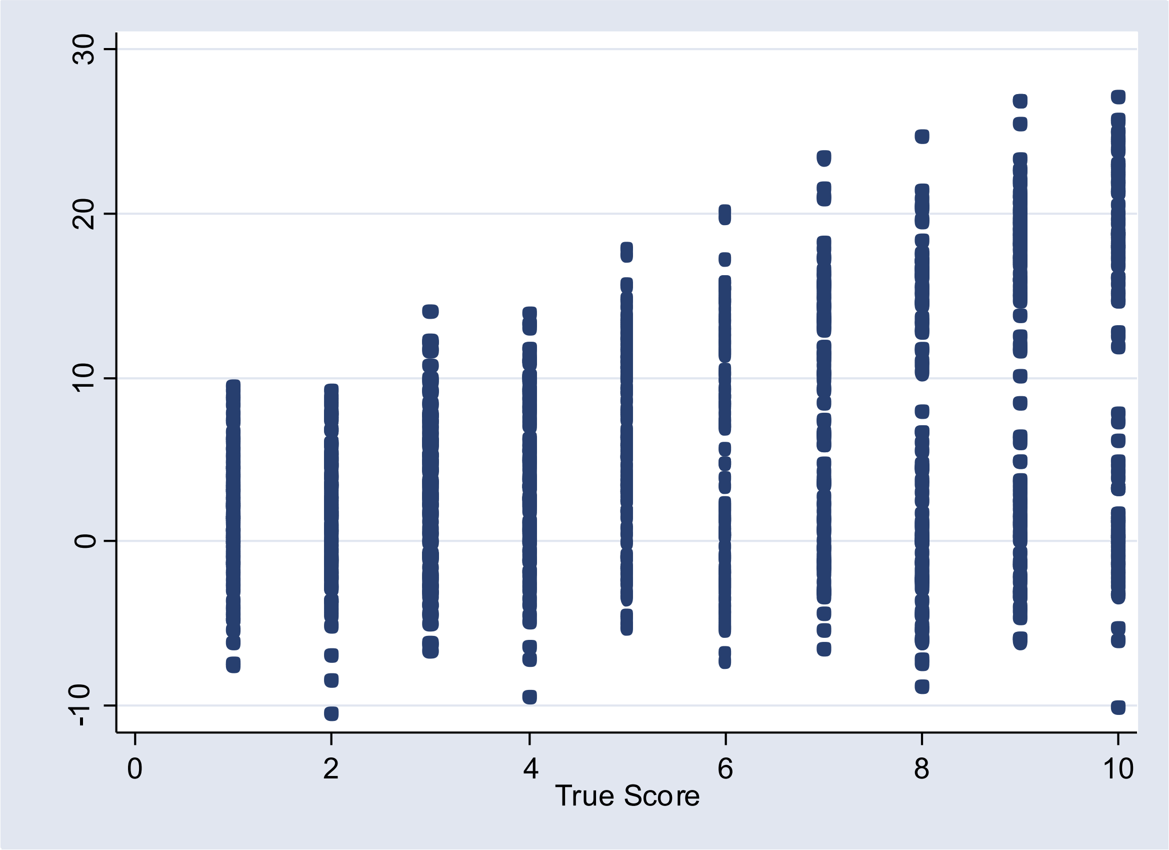
- A: Draw data where the cov(T_x,e) is negative
- B: Draw data where the cov(T_x,e) is positive



Exercise #2a – Answer



Exercise #2b - Answer



Exercise #3

- The reported correlations between years of educational attainment and adults’ scores on anti-social personality disorder scales (ASP) is usually about 0.30, and the reported reliability of the education scale is 0.95 and for the ASP scale 0.70. What will your observed correlation between these two measures be if your data on the education scale has the same reliability (0.95) but the ASP has much lower reliability of 0.40?

Exercise #3 - Answer

- Solve for true score correlation from reported data.

$$r_{TxTy} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} = \frac{.30}{\sqrt{.95 \times .70}} = .367883$$

- Solve for new observed correlation

$$r_{xy} = r_{TxTy} \times \sqrt{r_{xx} r_{yy}} = .367883 \times \sqrt{.95 * .40} = .227$$

Exercise #4

- In rating a dichotomous child health outcome among 100 children, two psychiatrists disagree in 20 cases – in 10 of these cases the 1st psychiatrist rated the outcome as present and the 2nd as absent, and in the other 10 cases were vice-versa. What will be the value of the Kappa coefficient if both psychiatrists agree that 50 children have the outcome?

Exercise #4 - Answer

Observed

	+	-	
+	50	10	60
-	10	30	40
	60	40	100

Proportion of observed agreement $= \frac{80}{100} = .8$

$$\kappa = \frac{p_{ob} - p_{ex}}{1 - p_{ex}} = \frac{.8 - .52}{1 - .52} = .58$$

Expected

	+	-	
+	60*60/100=36	60*40/100=24	60
-	60*40/100=24	40*40/100=16	40
	60	40	100

Proportion of expected agreement $= \frac{36+16}{100} = .52$

Exercise #5

- Give substantive examples of how measures of self-reported discrimination could possibly violate each of the three assumptions of classical test theory.

Exercise #5 - Answer

- $E(x) = 0$ could be violated if the true score is underreported as a result of social desirability bias
- $Cov(T_x, e) = 0$ could be violated if people systematically overreported or underreported discrimination at either high or low extremes of the measure
- $Cov(e_i, e_j) = 0$ could be violated if discrimination was clustered within certain areas of a location, and multiple locations were included in the analysis pool.