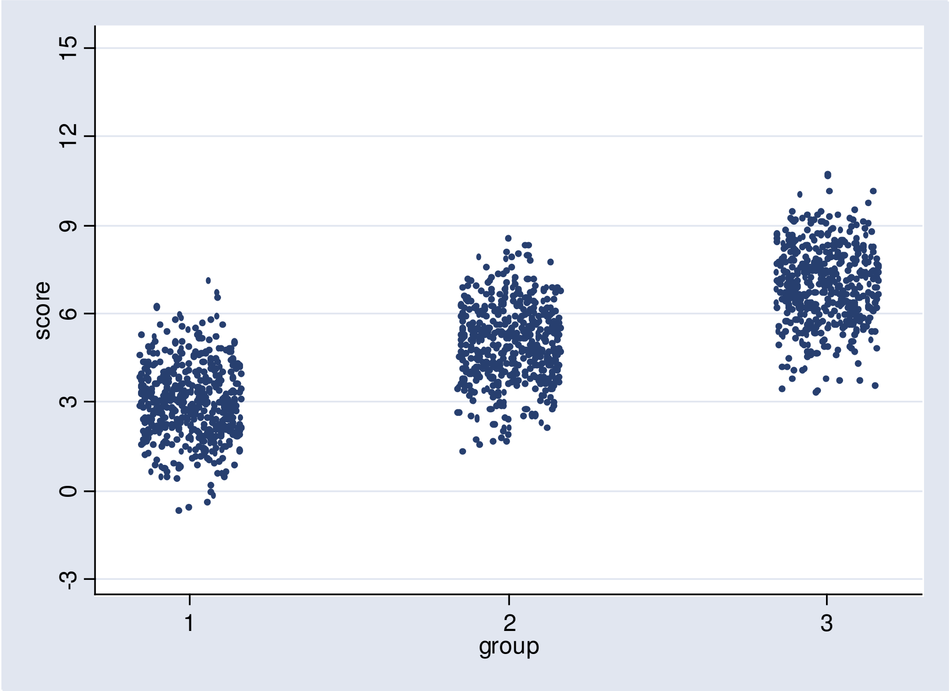
Reliability II

Lecture 4

9/13/06

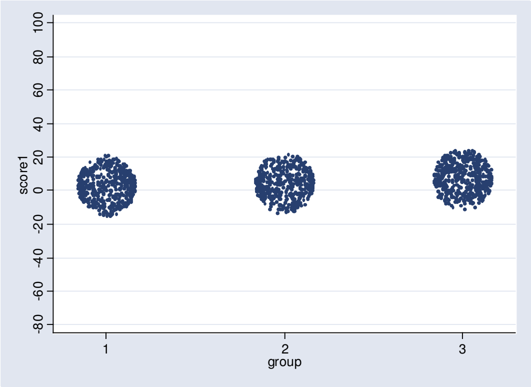
Outline

- Review of ANOVA
- Intra-Class Correlations
- Reliability Examples
- Other Research Designs



Question: are the true means for each group different from each other?

Compare amounts of variance within & between groups

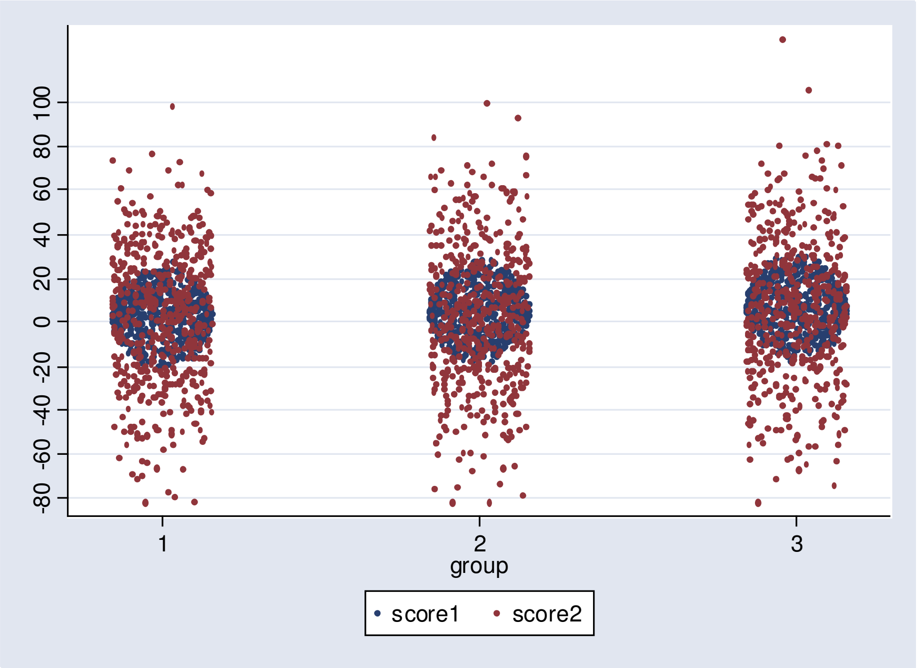


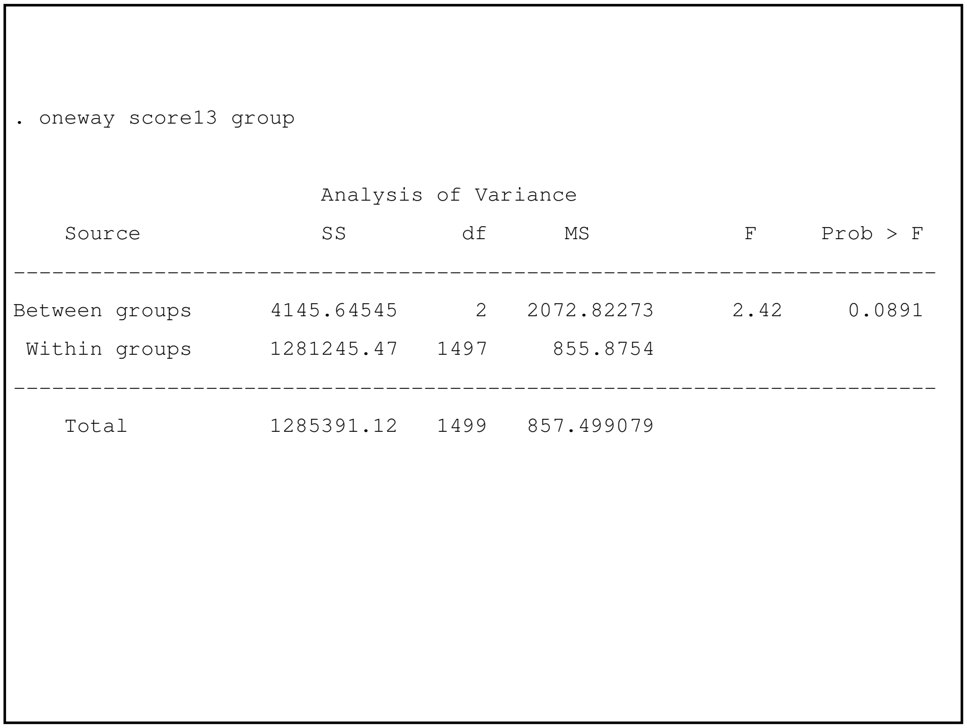
i=1…,I indexes groups, *j=1,…n*$_i$ indexes individuals

Source of Variation	DF	Sum of Squares (SS)	Mean Square (MS)	F-Ratio
Between	$I-1$	$\sum n_i(\overline{Y}_i-\overline{Y})^2$	$MSB=\dfrac{SSB}{DF}$	$\dfrac{MSB}{MSW}$
Within	$N-I$	$\sum_i\sum_j(Y_{ij}-\overline{Y}_i)^2$	$MSW=\dfrac{SSW}{DF}$	

```
. oneway score1 group

                        Analysis of Variance
    Source              SS         df      MS            F      Prob > F
-------------------------------------------------------------------------
Between groups      4054.76741      2   2027.38371    2030.92     0.0000
 Within groups      1494.39042   1497   .998256793
-------------------------------------------------------------------------
    Total           5549.15783   1499   3.70190649
```





```
. oneway score13 group

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      4145.64545      2   2072.82273      2.42     0.0891
 Within groups      1281245.47   1497    855.8754
------------------------------------------------------------------------
    Total           1285391.12   1499   857.499079
```

Intraclass correlation: Assessing inter-rater reliability

1) As before, reliability defined as:

$$\frac{\text{variance in true scores}}{\text{variance in the observed scores}}$$

2) For the intra-class correlation the specific form of this equation can take on at least six different forms

3) The correct form to use depends on the study design and the researcher's assumptions about the patients and subjects (or items)

4) I will discuss three designs, each with two ICCs

Overview: (raters might be people or questionnaire items)

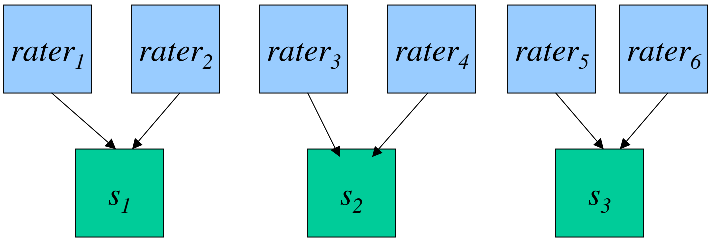
1. Unique Design:
   Each of the *I* subjects rated by a unique set of *m* raters *(m*>1), such that the total number of raters, *R,* is *m*I*

2. Fixed Design:
   Each subject is rated by each of the same *m* raters, such that the total number of raters, *R* is *m.* These raters are the <u>only</u> raters of interest.

3. Random Design:
   m raters are drawn from a larger pool of raters. Each of the *I* subjects is rated by each of the *m* raters. Again, the total number of raters, *R* is *m.*

Unique Design

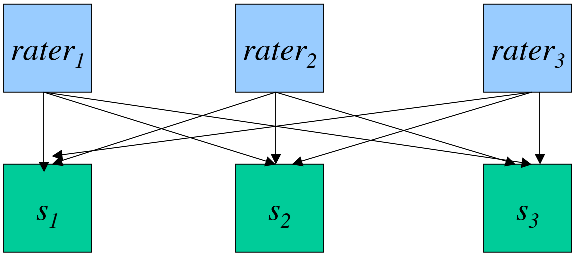
- No Overlap of Raters



- m=2, I=3 # of raters=m*I=6

Fixed Design

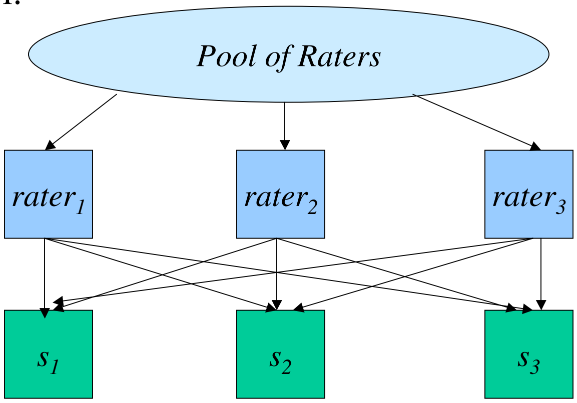
- Total Overlap of Raters



- m=3, n=3 # of raters=m=3

Random Design

- Total Overlap of Raters, but raters drawn from a pool.

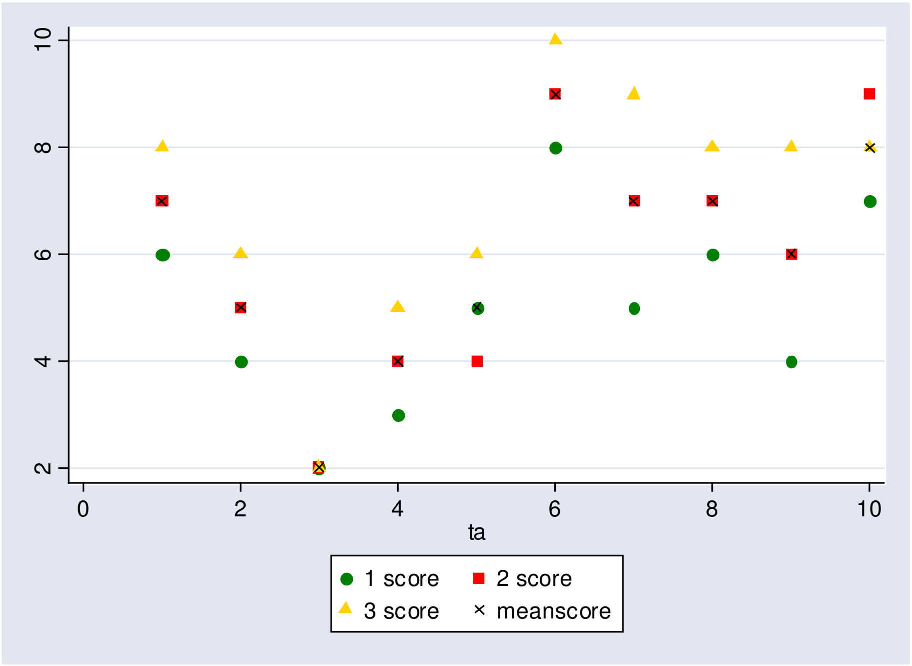


There are two (at least) types of reliability associated with each of these designs.

1) Reliability of mean ratings
reliability of average of all ratings per subject

2) Reliability of one individual rating
reliability of a single rating of one subject

3) Which will be higher? Why?



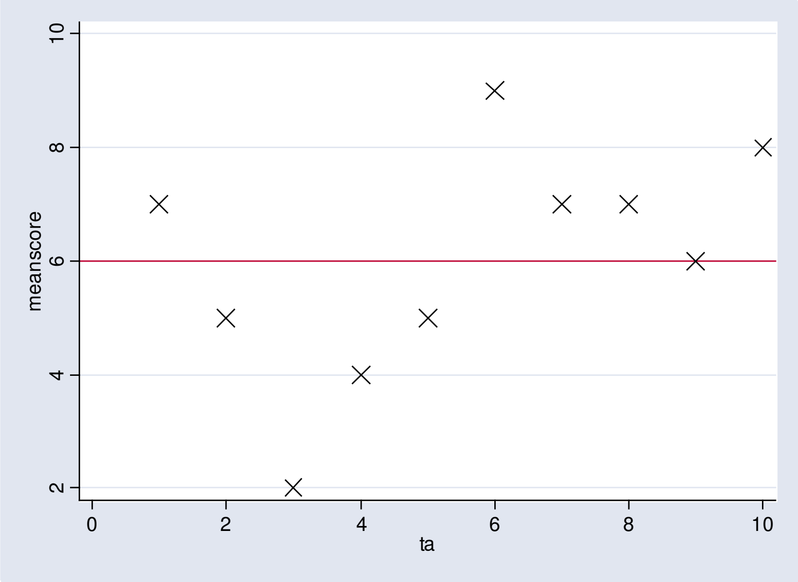
Unique Rater Design ICC:

Equation to estimate reliability of rating means:

$$\frac{\textit{Between Mean Square Variance} - \textit{Within Mean Square Variance}}{\textit{Between Mean Square Variance}}$$

$$\frac{MSB - MSW}{MSB}$$

Between Mean Score Variance (Each TA is a group): Observed mean variance



Between Mean Score Variance: Degree to which mean score of subjects being rated differ from grand mean

$$s_b^2 = \frac{1}{I-1}(m)\sum_{i=1}^{I}\left(\overline{Y}_i - \overline{Y}\right)^2 \cong \sigma_b^2$$

Where:

- I = number of people being rated (# of TAs)
- $\overline{Y}_i$ = mean score for each TA rated
- $\overline{Y}$ = overall mean of scores for whole sample
- m = number of raters for each mean

Unique Rater Design

1) Between Mean Score Variance, steps in Stata
   - a) calculate mean scores for each individual
     - *) e.g. egen meanta=rmean(score1 score2 score3)
   - b) calculate overall mean
     - *) e.g. egen grandmean=mean(meanta)
   - c) calculate deviation of individual mean from group mean
     - *) e.g. gen bsquarei=3*(meanta-grandmean)^2
   - d) add up all deviations in (c)
     - *) e.g. egen bsquare=sum(bsquarei)
   - e) divide sum of squares by degrees of freedom
     - *) e.g. display bsquare/(10-1) =

Unique Rater Design

2) Within Mean Score Variance: Degree to which individual scores differ from a subject's mean score

$$s_w^2 = \frac{1}{I\left(m-1\right)} \sum_{i=1}^{I} \sum_{j=1}^{m} \left(Y_{ij} - \overline{Y}_i\right)^2 \cong \sigma_w^2$$

Where:

- I = number of individuals being rated (# of TAs)
- R = number of raters
- Y_{ij} = score of each individual rater
- $\overline{Y}_i$ = mean score of each person rated
- m = number of raters for each mean

Note: *R=m*

Unique Rater Design

2) Within Mean Score Variance, steps in Stata

a) calculate mean scores for each individual

*) e.g. egen meanid=rmean(score1 score2 score3)

b) calculate deviation of rater from individual mean

*) e.g. gen wsquarei=

(score1-meanid)^2 + (score2-meanid)^2 +(score3- meanid)^2

c) add up deviations in (b) across all individuals

*) e.g. egen wsquare=sum(wsquarei)

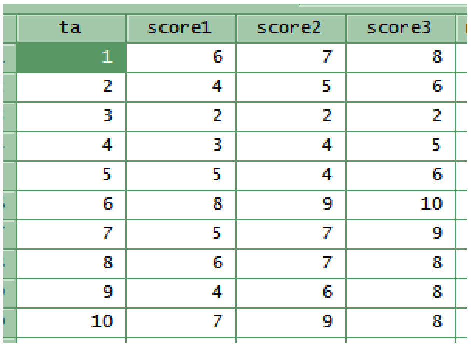
d) divide sum of squares by degrees of freedom

) e.g. display wsquare/I(m-1) =

Unique Rater Design

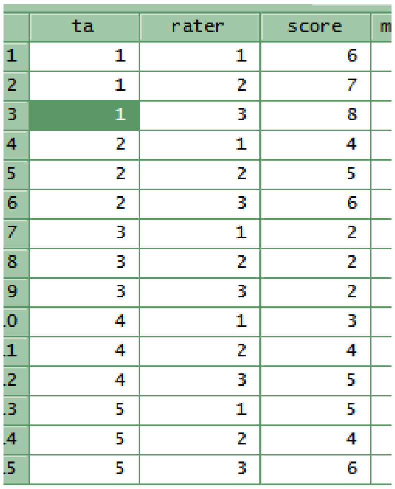
Shortcut: Use procedure ‘oneway’ in Stata

First, must “reshape” data.



ta	score1	score2	score3
1	6	7	8
2	4	5	6
3	2	2	2
4	3	4	5
5	5	4	6
6	8	9	10
7	5	7	9
8	6	7	8
9	4	6	8
10	7	9	8

```
. reshape long score, i(ta) j(rater)
```



	ta	rater	score	m
1	1	1	6	
2	1	2	7	
3	1	3	8	
4	2	1	4	
5	2	2	5	
6	2	3	6	
7	3	1	2	
8	3	2	2	
9	3	3	2	
0	4	1	3	
1	4	2	4	
2	4	3	5	
3	5	1	5	
4	5	2	4	
5	5	3	6	

Using ANOVA in STATA to calculate variance:

Example:

```
. oneway score ta

                        Analysis of Variance
    Source              SS         df      MS
------------------------------------------------------
Between groups          114.00      9   12.6666667
 Within groups           30.00     20         1.50
------------------------------------------------------
    Total               144.00     29   4.96551724
```

$$ICC = \frac{MSB - MSW}{MSB}$$

$$ICC = \frac{12.67 - 1.50}{12.67} = .8816$$

Important note:

Reliability is a group-specific statistic.

The greater the variance in the true scores of a population, the higher the reliability of the measure (if observed variance is constant)

$$\text{Reliability} = \frac{\text{variance in true scores}}{\text{variance in observed scores}}$$

Reliability for individual ratings

So far we've calculated reliability of the mean score for each TA.

What is the average reliability of each individual rating of the TA?

Reliability of Individual Scores in Unique Rater Design:

Equation:

$$\frac{MSB - MSW}{MSB + (m-1)\ MSW}$$

Where m = number of raters per TA

Continuing with our example:

$$\text{Re}\,liability = \frac{(12.67 - 1.50)}{12.67 + (3-1) * 1.50} = .7128$$

Fixed Rater Design

1) Each subject rated by each of the same m raters, who are the only raters of interest

2) examples:

3) Computation involves two-way analysis of variance

4) Before: two sources of error, (differences across individuals, and error inherent to the measurement) Error now only has one source: error due to individuals is 'controlled.'

Fixed Rater Design

Recall that the equation for Unique Rater Design was:

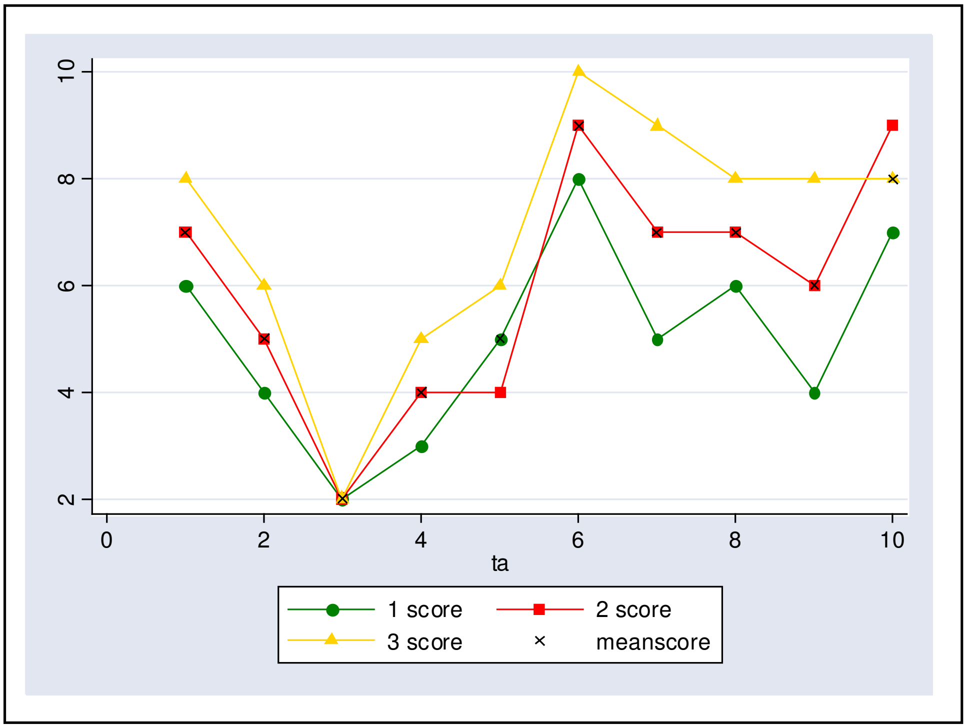
$$\frac{MSB - MSW}{MSB}$$

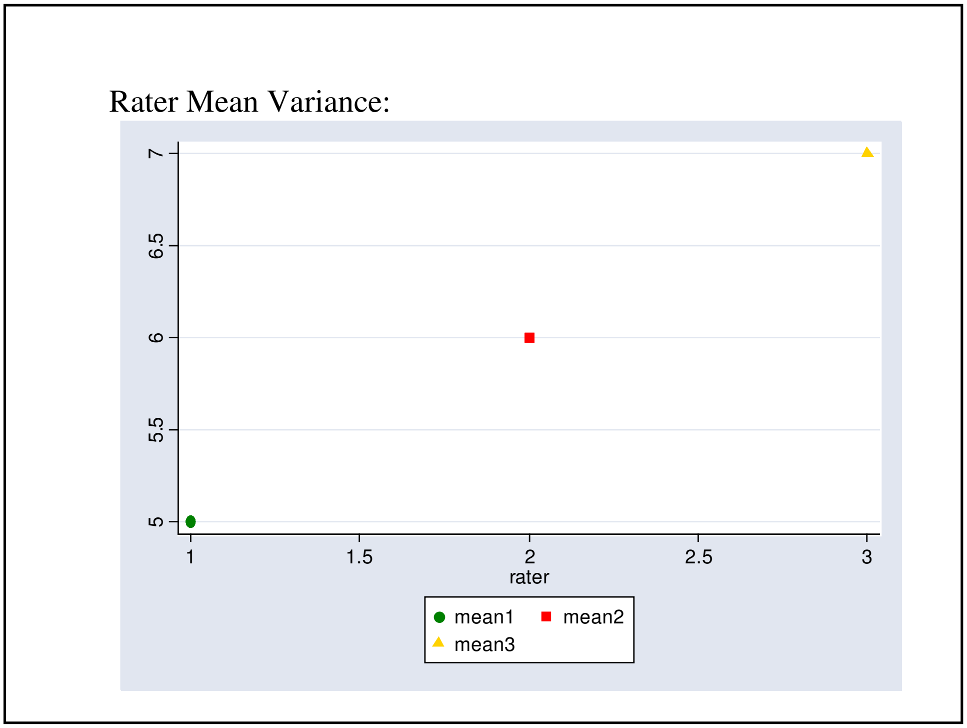
Which can also be expressed as:

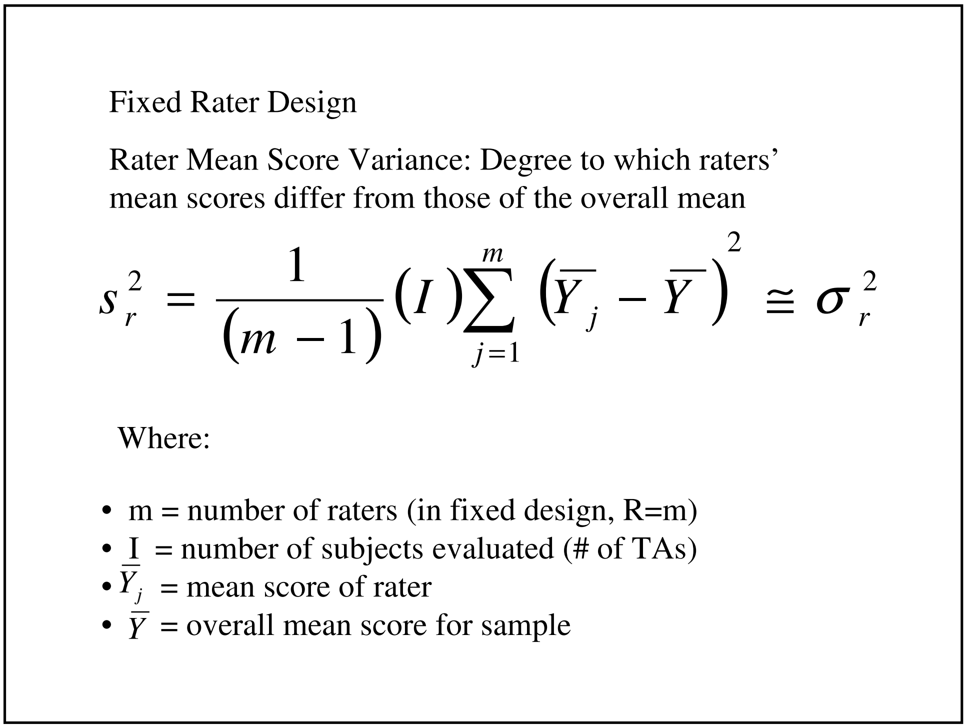
$$\frac{MSB - (MSRater + MSE)}{MSB}$$

The equation for the fixed rater design is very similar:

$$\frac{MSB - (MSE)}{MSB}$$




Rater Mean Variance:



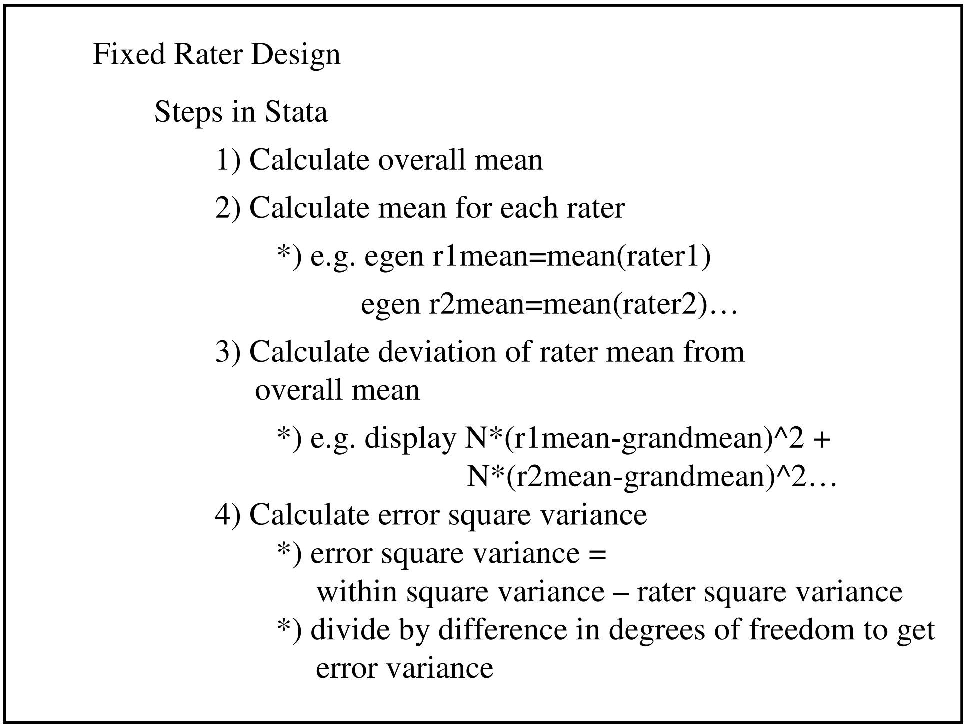
Fixed Rater Design

Rater Mean Score Variance: Degree to which raters' mean scores differ from those of the overall mean

$$s_r^2 = \frac{1}{(m-1)}(I)\sum_{j=1}^{m}\left(\overline{Y}_j - \overline{Y}\right)^2 \cong \sigma_r^2$$

Where:

- m = number of raters (in fixed design, R=m)
- I = number of subjects evaluated (# of TAs)
- $\overline{Y}_j$ = mean score of rater
- $\overline{Y}$ = overall mean score for sample



Fixed Rater Design

Steps in Stata

1) Calculate overall mean

2) Calculate mean for each rater

*) e.g. egen r1mean=mean(rater1)

```
egen r2mean=mean(rater2)…
```

3) Calculate deviation of rater mean from overall mean

) e.g. display N(r1mean-grandmean)^2 +
N*(r2mean-grandmean)^2…

4) Calculate error square variance

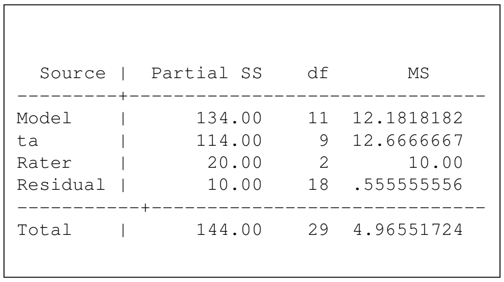
*) error square variance =
within square variance – rater square variance

*) divide by difference in degrees of freedom to get error variance

Using ANOVA in STATA to calculate variance:

Example:

```
. anova score ta observer
```



```
  Source |  Partial SS    df       MS
---------+-------------------------------
Model    |      134.00    11  12.1818182
ta       |      114.00     9  12.6666667
Rater    |       20.00     2       10.00
Residual |       10.00    18  .555555556
-----------+-----------------------------
Total    |      144.00    29  4.96551724
```

ICC for Fixed Rater Design, Group mean =

$$\frac{MSB - MSE}{MSB} = \frac{12.67 - .56}{12.67} = .96$$

Fixed Rater Design

Equation to estimate reliability for individual rater's scores:

$$\frac{\textit{Between Mean Square Variance – Error Mean Square Variance}}{\textit{Between Mean Square Variance + (m-1)*Error Mean Square}}$$

Where R= m=number of raters

Final Estimate:

$$\frac{12.67 - .56}{12.67 + (2)(.56)} = .8782$$

Random Rater Design

1) Randomly-selected raters evaluate each subject

2) Computation involves two-way analysis of variance

3) Error has two sources again, but error due to individual raters is reduced

4) Deciding between Random and Fixed design:

Would you wish to generalize findings from this sample to situations with a different set of raters? If so, you would use the random rater design.

Random Rater Design 1) Reliability for mean score of each subject:

$$\frac{MSB - MSE}{MSB + ((MSRater - MSE)/I)}$$

```
Source |  Partial SS    df       MS
---------+--------------------------------
Model    |      134.00    11  12.1818182
ta       |      114.00     9  12.6666667
Rater    |       20.00     2       10.00
Residual |       10.00    18  .555555556
-----------+------------------------------
Total    |      144.00    29  4.96551724
```

2) Take into account error for rater bias

3) ICC = $\frac{12.67 - 0.56}{12.67 + (10 - 0.56)/10}$ = .89

Random Rater Design

1) Reliability for individual scores:

```
Source |  Partial SS    df       MS
---------+-------------------------------
Model    |      134.00    11  12.1818182
ta       |      114.00     9  12.6666667
Rater    |       20.00     2       10.00
Residual |       10.00    18  .555555556
-----------+-----------------------------
Total    |      144.00    29  4.96551724
```

$$\frac{MSB - MSE}{MSB + (m-1) * MSE + m * (MSRater - MSE) / I}$$

2) ICC = $\dfrac{12.67 - 0.56}{12.67 + (2 * .56) + 3 * (10.0 - .56) / 10}$ = .72

Summary:

1) Unique Rater Design: Each subject rated by a different set of *m* raters

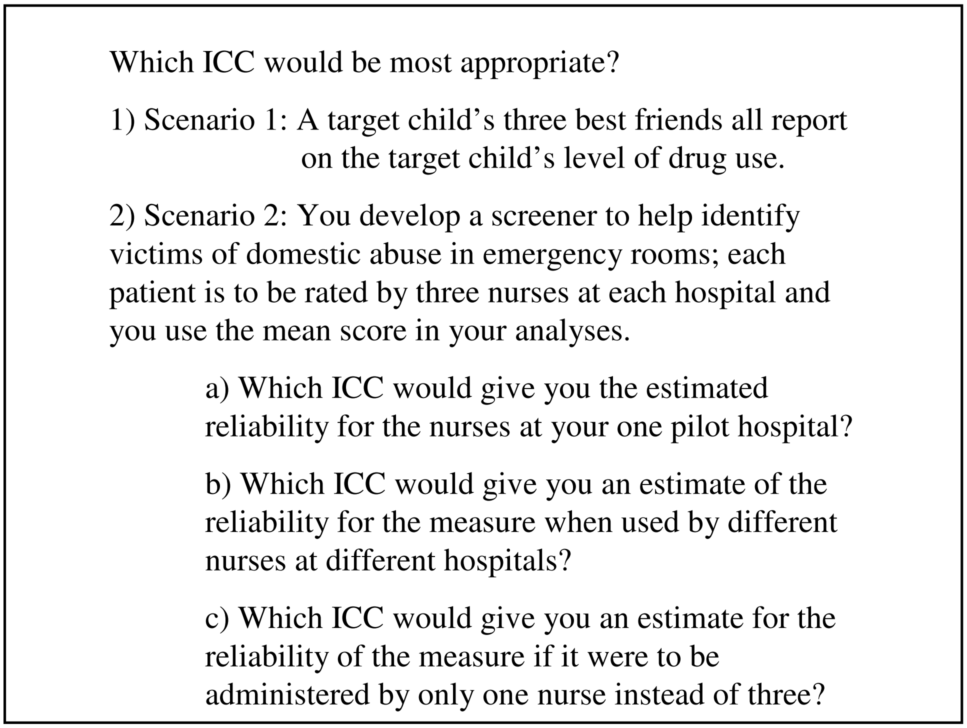
a) formulas use between and within mean square variance

2) Fixed Rater Design: Each target is rated by each of the same *m* raters, who are the only raters of interest

a) formulas use between and error square variance

3) Random Rater Design: *m* raters, in (2), were drawn from a random sample of raters

a) formula uses between and error square variance, adjusting for rater variance



Which ICC would be most appropriate?

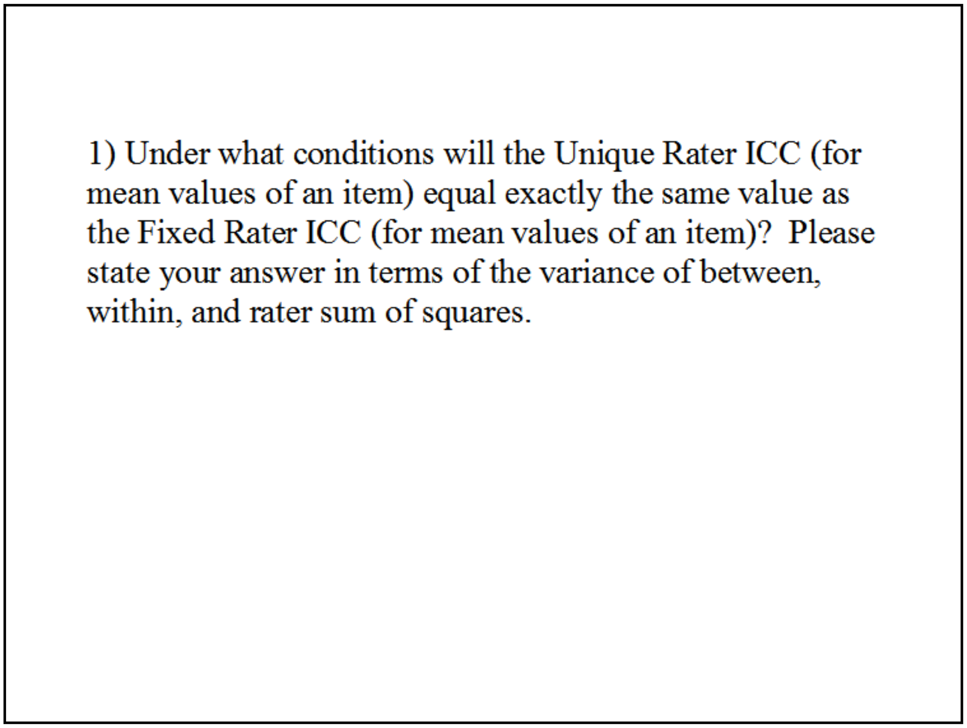
1) Scenario 1: A target child’s three best friends all report on the target child’s level of drug use.

2) Scenario 2: You develop a screener to help identify victims of domestic abuse in emergency rooms; each patient is to be rated by three nurses at each hospital and you use the mean score in your analyses.

a) Which ICC would give you the estimated reliability for the nurses at your one pilot hospital?

b) Which ICC would give you an estimate of the reliability for the measure when used by different nurses at different hospitals?

c) Which ICC would give you an estimate for the reliability of the measure if it were to be administered by only one nurse instead of three?

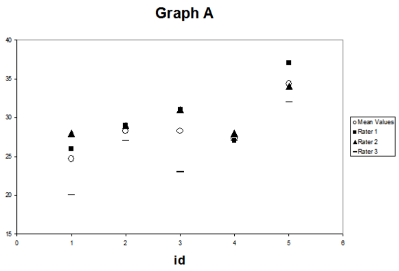


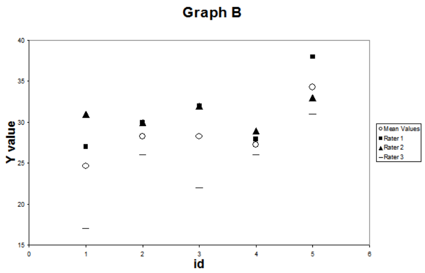
1) Under what conditions will the Unique Rater ICC (for mean values of an item) equal exactly the same value as the Fixed Rater ICC (for mean values of an item)? Please state your answer in terms of the variance of between, within, and rater sum of squares.

2) You develop a new survey measure of depression based on a pilot sample that consists 33% of people with severe depression, 33% of people with mild depression, and 33% of people without any depression. You are happy to discover that your measure has a high reliability of .90 (which is very high for such a measure!). 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?

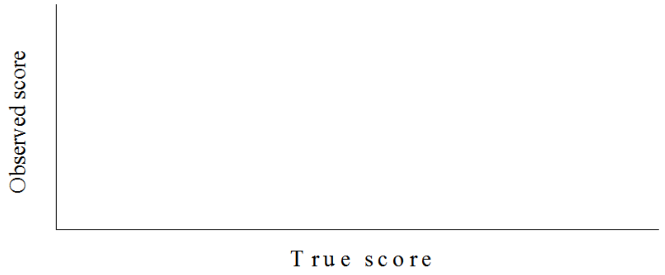
3) Steve says, “I’m a little confused. Intuitively, high reliability means that if you measure the same characteristic twice you should get the same answer. But in class the professor drew graphs that seem to imply that reliability will be higher when the variability in the sample is higher.” What is your response to Steve?

4) Which measure has a higher reliability, (a) or (b)? Why?

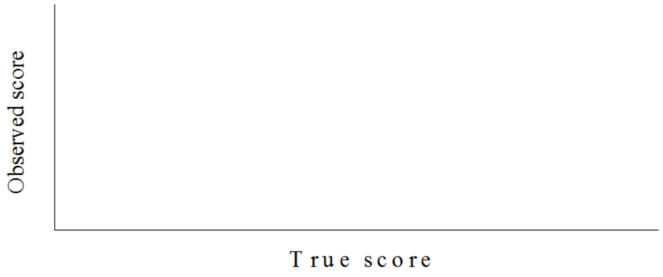


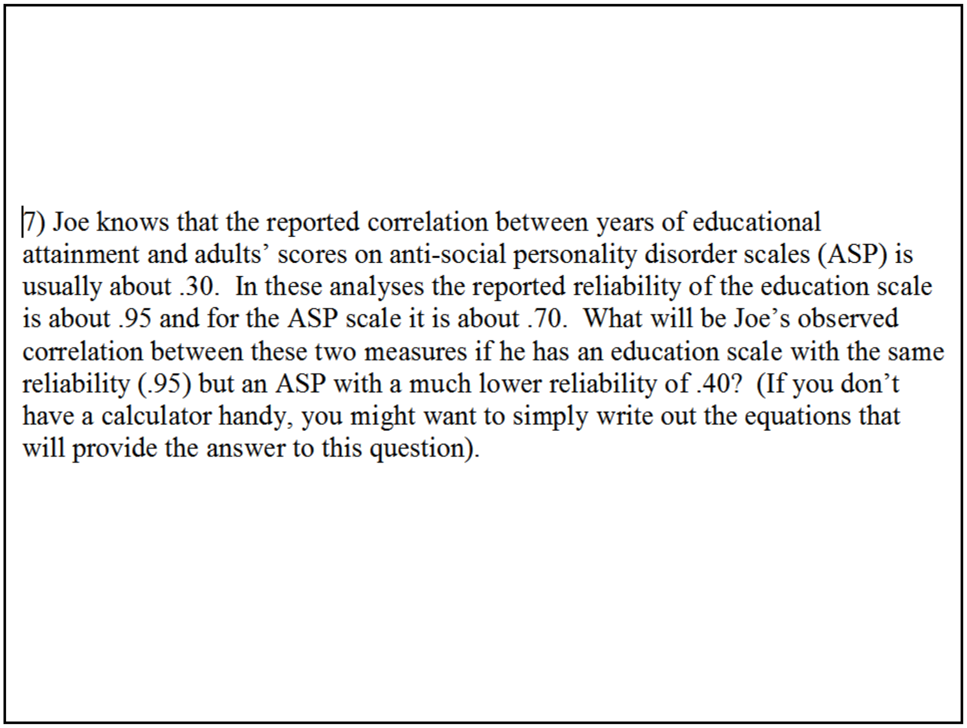


5) Draw an example of a measure that has a negative covariance between the true score and the error term.



6) Draw an example of a measure that has a positive covariance between the true score and the error term.





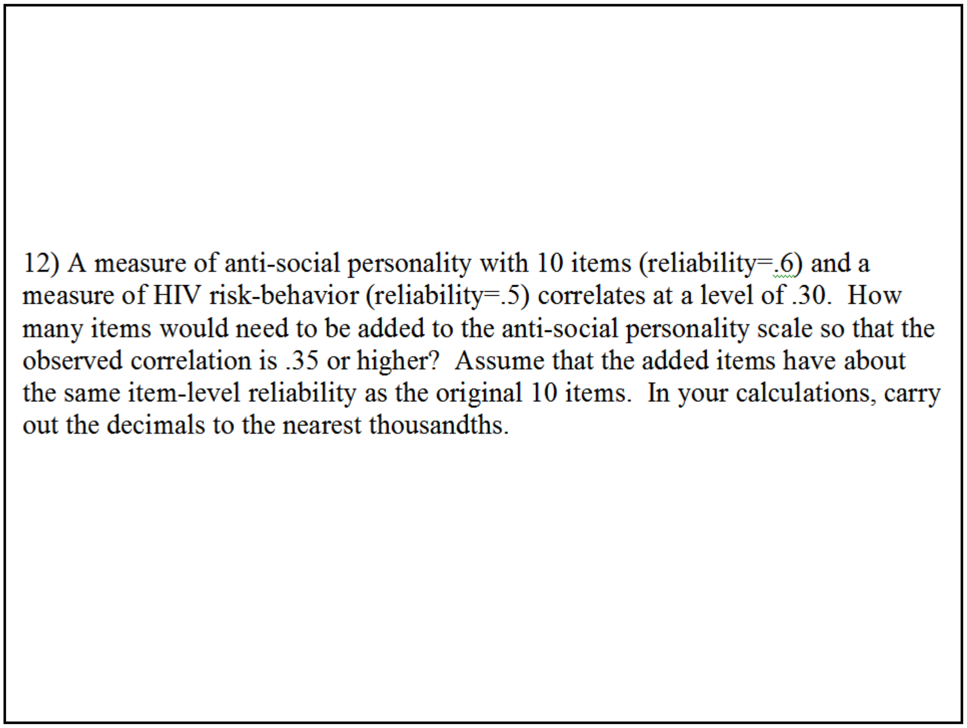
7) Joe knows that the reported correlation between years of educational attainment and adults' scores on anti-social personality disorder scales (ASP) is usually about .30. In these analyses the reported reliability of the education scale is about .95 and for the ASP scale it is about .70. What will be Joe's observed correlation between these two measures if he has an education scale with the same reliability (.95) but an ASP with a much lower reliability of .40? (If you don't have a calculator handy, you might want to simply write out the equations that will provide the answer to this question).

8) How, conceptually, is the alpha related to the split-half reliability coefficient? How is the alpha related to the Fixed Rater ICC for mean scores?

9) If the reliability for a ten-item scale with an average inter-item correlation of .25 is .75, what would be the reliability of a twenty-item scale with the same average inter-item correlation? What would be the reliability of a 15-item scale? Of a 5 item scale?

10) 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? Will the Kappa be higher or lower if they agree that 70 children have the outcome?

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

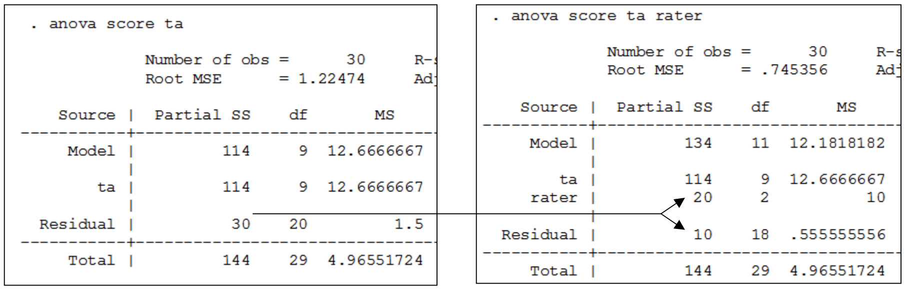


12) A measure of anti-social personality with 10 items (reliability=.6) and a measure of HIV risk-behavior (reliability=.5) correlates at a level of .30. How many items would need to be added to the anti-social personality scale so that the observed correlation is .35 or higher? Assume that the added items have about the same item-level reliability as the original 10 items. In your calculations, carry out the decimals to the nearest thousandths.

13) Give examples of how children's self-report of depression could be reliable but not valid.

Other Research Designs

- We saw, with the fixed ICC, how we could partition the variance, and reduce MSE



```
. anova score ta

                 Number of obs =      30     R-
                 Root MSE      = 1.22474     Adj

    Source |  Partial SS    df       MS
-----------+----------------------------------
     Model |         114     9  12.6666667
           |
        ta |         114     9  12.6666667
           |
  Residual |          30    20         1.5
-----------+----------------------------------
     Total |         144    29  4.96551724
```

```
. anova score ta rater

                 Number of obs =      30     R-
                 Root MSE      = .745356     Adj

    Source |  Partial SS    df       MS
-----------+----------------------------------
     Model |         134    11  12.1818182
           |
        ta |         114     9  12.6666667
     rater |          20     2          10
           |
  Residual |          10    18  .555555556
-----------+----------------------------------
     Total |         144    29  4.96551724
```

Fixed Effects

(a) Set by experimenter (eg, treatment in an RCT)

(b) it is unreasonable to generalize beyond conditions. (eg, reading ability as a function of grade in school)

(c) when the # of possibilities is small, and all are included in the study design (eg, sex, in a study with both males and females)

Random Effects

(a) Multiple possible values (eg, personality measures, age).

(b) Study subjects are considered a representative sample from a larger population.

(c) Experimenter wishes to *generalize* the results of the study beyond the study sample.

- We already saw an example of this with the fixed and random ICC's.
- Part of a larger group of study designs under the heading of "generalizability theory" popularized by Cronbach, and others.
- Can take 140.655 (LDA) and/or 140.656 (Multilevel models)