Lecture 1.2: Reliability of quantitative scores

- **Task**: Establish the reliability of novel index, *X*, to be used in substantive analyses, the quality of which depends on quality of *X*. The toolkit:
- If *X* is *categorical*, reliability can be assessed by **Cohen's *kappa*** (κ), as in Lec. 1.1.
- If *X* is *quantitative*, reliability can be assessed by **correlation**, **Cronbach's *alpha*** (α), ***ICC*** [with `ICC(matrix)` in the **psych** package], etc., as will be seen.
- If *X* is *ordinal*, reliability can be assessed by a **weighted *kappa***, as in Lec. 1.1.

Lecture 1.2: Reliability of quantitative scores

Index		Reliability Measure
Categorical Agreement between raters (e.g., “yes”/”no”)	→	Cohen’s κ
Quantitative Relative agreement/consistency (e.g., 1-5)	→	Correlation, Cronbach’s α, ICC
Ordinal Weighted agreement (e.g., “absent”, “mild”, “severe”)	→	Weighted κ

Heuristics used to compute reliability

- *Reliab* as ‘**agreement**’ between raters (e.g., Cohen’s κ for categorical scores)
- *Reliab* as ‘**weighted agreement**’, when scores are ordinal and some disagreements count more than others
- *Reliab* as ‘relative agreement’, e.g., agreement in the **rankings** of objects; indexed by **correlation, *r***
- *Reliab* as ‘consistency’ or ‘precision’, indexed, e.g., by **proportion of variance accounted for** by objects relative to ‘error’; sometimes **=** r^2**, for some ‘*r*’**

Meta-theoretical issues

- At what level of analysis should we conduct a reliability analysis?
  - In HW-1: the ‘thought’ or the ‘participant’?
  - In McLoyd et al (Lec 1.1): the ‘30 sec interval’ or the ‘30 min exptl session’?
- Are ratings categorical, ordinal or interval-scaled (quantitative)? How many raters? Are some ‘experts’ and some ‘novices’? What is the data format? The answers influence *reliability?*
- Is a single index, e.g., *κ,* sufficient or satisfactory, or shd we use a cognitive model to interpret the raters’ scores? (To be discussed)

Lecture Outline

- {**Data** Format} → {Formula for computing *reliability* of *X*}, and we should understand the **theory** justifying the formula used in computation. E.g., the χ^2 approach used for Cohen's κ; the ANOVA approach for *ICC* and Cronbach's α. We consider a few 'formats' and 'heuristics' to illustrate the range of formulae.
- Some **R examples** if there's time
- **Intro to Test Theory:** How to design optimal 'tests' (or 'scales'), e.g., by using 'consistent' items, or 'long' tests? (See HW-1, #4)

- **1-way design**: On each visit to a clinic, patients are rated by whomever is on duty. After 3 visits, the clinic supervisor examines **all 3 ratings** and determines the patient's final rating or 'clinical status'. We wish to determine the reliability of a patient's clinical status.

Patient	**Ratings**
A	3, 5, 1
C	4, 2, 5
D …	1, 3, 3

- But what is the reliability if future ratings were based **on 1 doctor**, instead of 3 doctors?

- **Weighted affective valence estimates (WAVEs) of colors** (Schloss & Palmer, 2009). Ask 74 Ps to list the **objects** they associate with each of **37 colors**. This generated 4000+ objects that were then reduced to **280** object classes.

Color	**P1**	**P2**
red	apple, firetruck	cherry, apple
blue	blueberry, box	ocean, berry
green	shamrock, grass	leaf, tree

…37 colors ....

- **Weighted affective valence estimates (WAVEs) of colors** (Schloss & Palmer, 2009). Ask 74 Ps to list the **objects** they associate with each of **37 colors**. This generated 4000+ objects that were then reduced to **280** object classes.

Object Class	**P1**	**P2**
fruit		
vehicle		
vegetation		

…280 classes

- **2-way design**. Next, 98 Ps rate how positive/ appealing **each** obj-class is, and these ratings are averaged for each obj-class. How reliable are the average ratings? (The burden on P is onerous!)

…98 Ps

Object Class	**P1**	**P2**	**Avg**
fruit	6	5	5.5
vehicle	2	4	3
vegetation	4	4	4

- **1-way design**. Instead, each of 98 Ps rates how positive/appealing each of **only 20** obj-classes is, and these ratings are averaged for each obj-class. How reliable are the average ratings?

…98 Ps

Object Class	P1	P2	Avg
fruit	6		6
vehicle		4	4
vegetation			4

- **To calculate the WAVE of each color:** Let p_i be the relative frequency with which obj-class is listed as associated with the color; and let v_i be the average valence of the *i*'th obj-class. Then the WAVE for that color is: $WAVE = \Sigma p_i v_i$.

Color	**WAVE**	
red	vehicle p = .1*4	fruit p = .7*6

- **To calculate the WAVE of each color:** Let p_i be the relative frequency with which obj-class is listed as associated with the color; and let v_i be the average valence of the *i*'th obj-class. Then the WAVE for that color is: $WAVE = \Sigma p_i v_i$.
- *WAVE* is a much **better predictor of 'color preference'** than a physiological theory based on opponent cone outputs and gender differences.

- **To calculate the WAVE of each color:** Let p_i be the relative frequency with which obj-class is listed as associated with the color; and let v_i be the average valence of the *i*'th obj-class. Then the WAVE for that color is: $WAVE = \Sigma p_i v_i$.
- *WAVE* is a much **better predictor of 'color preference'** than a physiological theory based on opponent cone outputs and gender differences.
- What is the **reliability** of *WAVE*? **Ans**. One possibility is to 'randomly divide the P's into 2 halves', and compute $WAVE_1$ and $WAVE_2$ for the 2 halves, and compute **corr**($WAVE_1$, $WAVE_2$) across the 37 colors. This corr can be used to get reliability ('split-half reliability').

- **A typical 2-way design**: Three (3) summer RA's rate the same set of 50 5-sec clips of dyadic interaction – for 'intimacy', etc. After satisfying ourselves that the ratings of the RA's are reliable or consistent, **each RA is given his or her own disc with clips to rate**.
- The ratings of the behavior on the discs are then analysed to test our substantive theory. We wish the reliability of the ratings on which our analysis is based, namely, the reliability of a single rater's score, **not** the average of 2 raters!

	D1	D2	D3
P1	3.9	4.7	5.0
P2	8.4	8.1	6.4
P3	7.6	7.6	5.8
P4	5.6	7.3	5.2
P5	1.1	2.7	3.5
P6	8.0	9.8	9.1
P7	8.7	6.1	7.0
P8	8.5	6.1	5.0
P9	5.2	7.6	5.4
P10	4.4	5.1	3.8
P11	0.6	3.3	1.1
P12	5.6	6.4	4.3
P13	8.8	9.3	8.2
P14	5.8	5.4	5.3
P15	6.1	5.5	7.9

(Data file in 'short' form; RA's are D1, D2, D3, Clips are in Rows as P1, …)

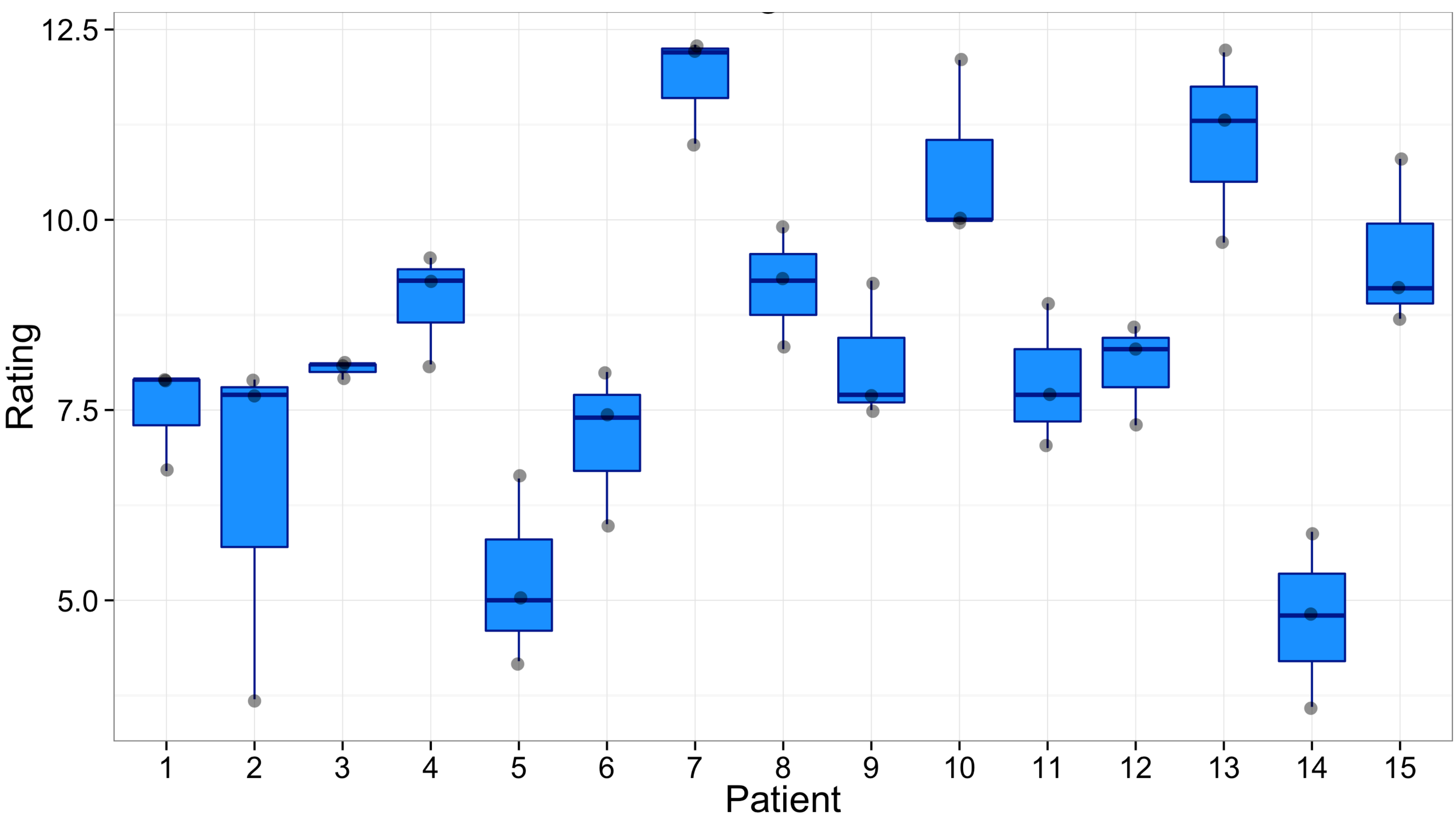
Summary for 6 different ‘designs’

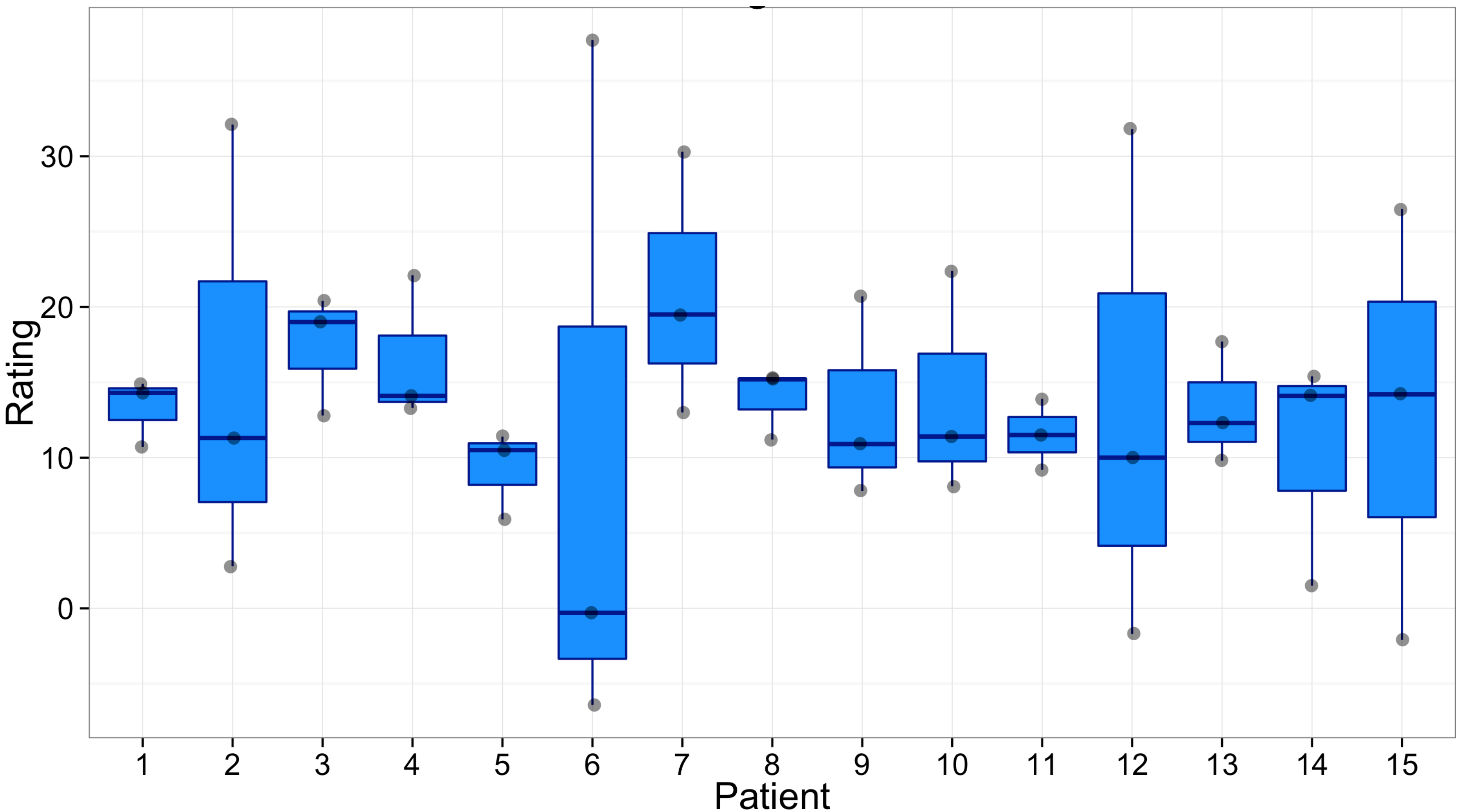
Assignment of *n* Targets to *k* Raters: For each assignment, is reliability to be assessed for a *single* rater or for the *average* of *k* raters?
1. Each Target is rated by a randomly chosen set of *k* Raters (1-way design)
2. A random sample of *k* Raters, each Rater rates all Targets (2-way design with Rater as random effect; so reliability estimate generalises to **popn** of Raters)
3. A fixed set of *k* Raters, each Rater rates all Targets (2-way design with Rater as fixed effect; NO generalisation)

Calculations and Theory

1. *ICC* for the 1-way design: Each of *p* objects is rated by *k* (= 3) randomly chosen *raters.* This is a 1-way design.

P or Object	Scores
1	5, 2, 6
2	2, 4, 4
3	0, 4, 2
4	.., .., ..


12.5
10.0
7.5
5.0
Rating
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Patient



Rating
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Patient

For more examples, check out: **http://stanford.edu/class/psych253/apps/**

Visualizing the Intraclass Correlation Coefficient (ICC)

Relate data variability to ICC derived from assorted models

Model to use

1-way random effects

New Sample

Generating parameters

Number of Raters

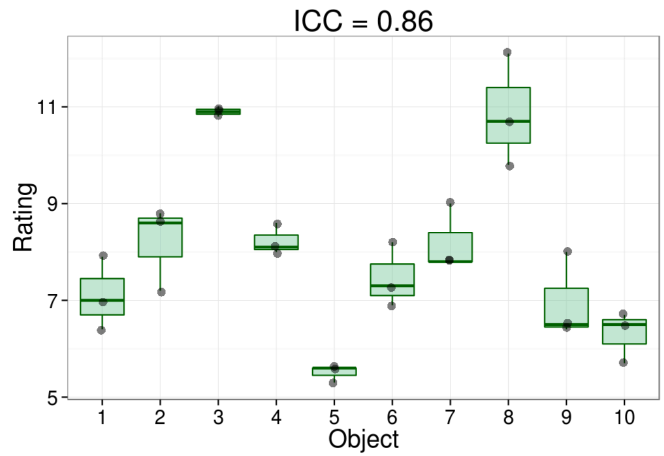
2 3

Number of Objects to Rate

5 10 30

Measurement Error

1 10




```
Call: ICC(x = data$d0)

Intraclass correlation coefficients
                         type  ICC  F df1 df2       p lower bound upper bound
Single_raters_absolute   ICC1 0.86 19   9  20 5.1e-08        0.66        0.96
Single_random_raters     ICC2 0.86 19   9  18 1.8e-07        0.66        0.96
Single_fixed_raters      ICC3 0.86 19   9  18 1.8e-07        0.65        0.96
Average_raters_absolute ICC1k 0.95 19   9  20 5.1e-08        0.85        0.99
Average_random_raters   ICC2k 0.95 19   9  18 1.8e-07        0.85        0.99
Average_fixed_raters    ICC3k 0.95 19   9  18 1.8e-07        0.85        0.99
```

ICC and ANOVA

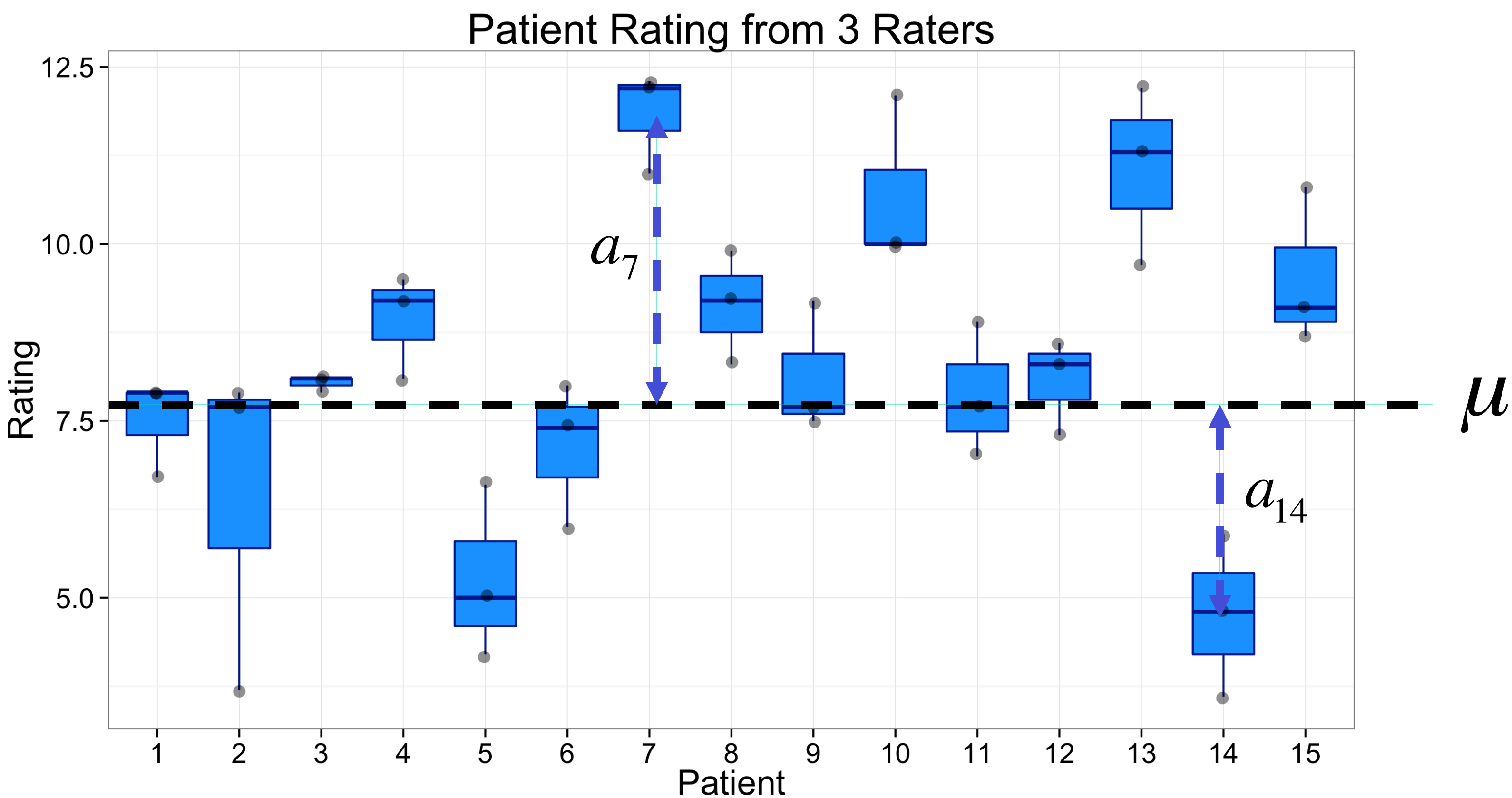
- To pursue the idea that ICC measures the **within-group similarity relative to between-group similarity**, consider the 1-way ANOVA model: *i* indexes the groups, and *j* indexes the *j*' th obs, Y_{ij}, within the *i*' th group: $Y_{ij} = \mu + a_i + e_{ij}$

- In the fixed effects ANOVA model, we treat the groups as 'fixed', e.g., male vs female, or low vs medium vs high. In the present context, it is more reasonable to treat the **groups as having been randomly selected**, e.g., objects, households, twins. Hence we use the '**random effects**' model of ANOVA.

$$Y_{ij} = \mu + a_i + e_{ij}$$

- Y_{ij} is the score of the *j*' th person (rater) in the *i*' th group (object), μ is the overall population mean.
- a_i is a random variable representing the 'effect of being in group *i*'. Everyone in group *i* has the same value of a_i, which varies randomly from group to group, with a mean of 0 and a variance of σ_a^2.

$$Y_{ij} = \mu + a_i + e_{ij}$$



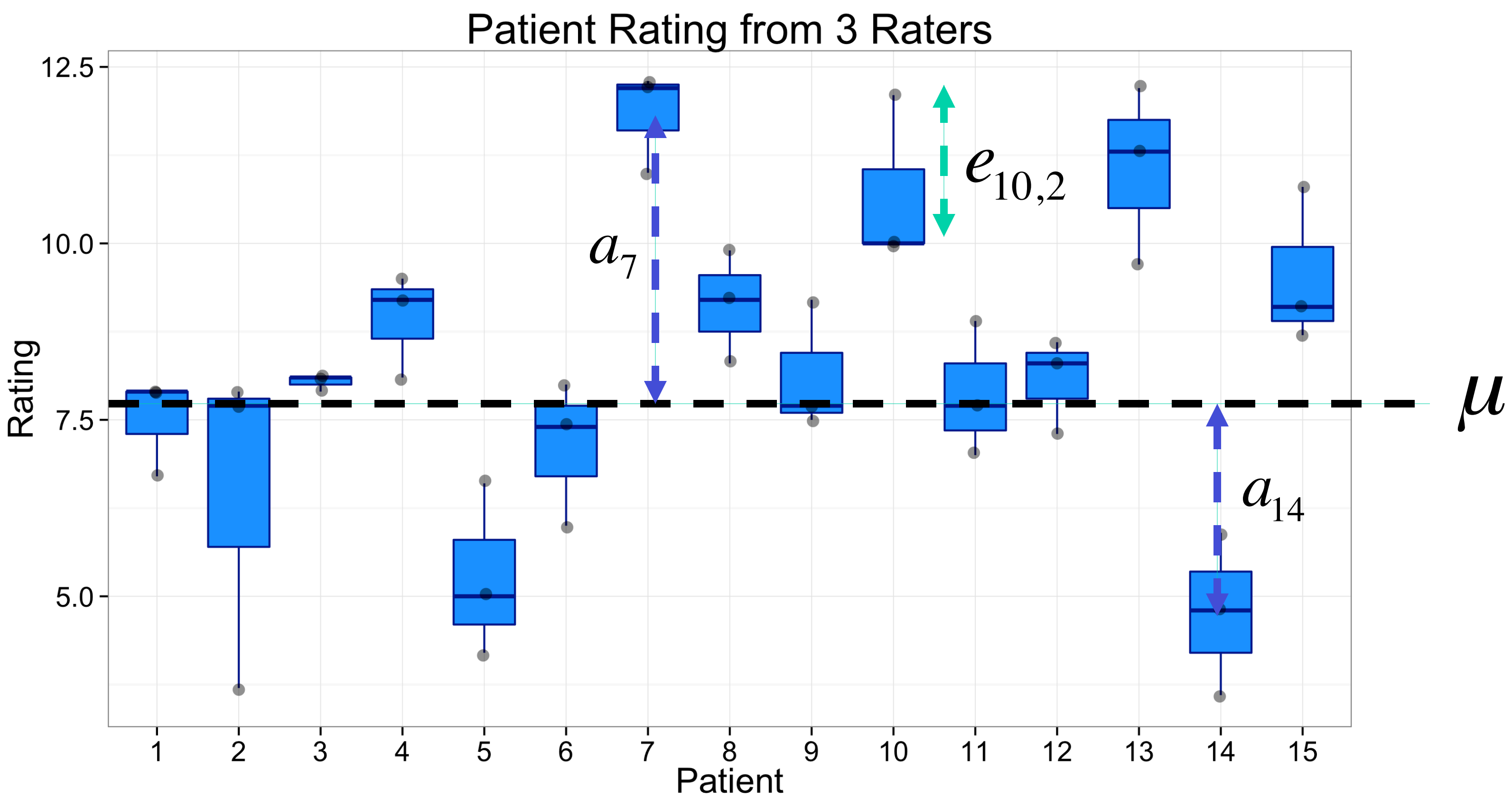
Patient Rating from 3 Raters
12.5
10.0
7.5
5.0
Rating
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Patient
a_7
a_{14}
μ

$$Y_{ij} = \mu + a_i + e_{ij}$$

- e_{ij} is a random variable representing the 'measurement error' in the score of the *j*' th person in group *i*. e_{ij} is a random variable with a mean of 0 and a variance of σ_e^2, and it is independent of a_i.
- The 'total variance' is the variance of Y_{ij}. This variance is the sum of the variances of a_i and e_{ij}:

  var(Y_{ij}) = var(a_i) + var(e_{ij}) = $\sigma_a^2 + \sigma_e^2$.
- The *ICC* is defined as the fraction of total variance accounted for by variation across groups.

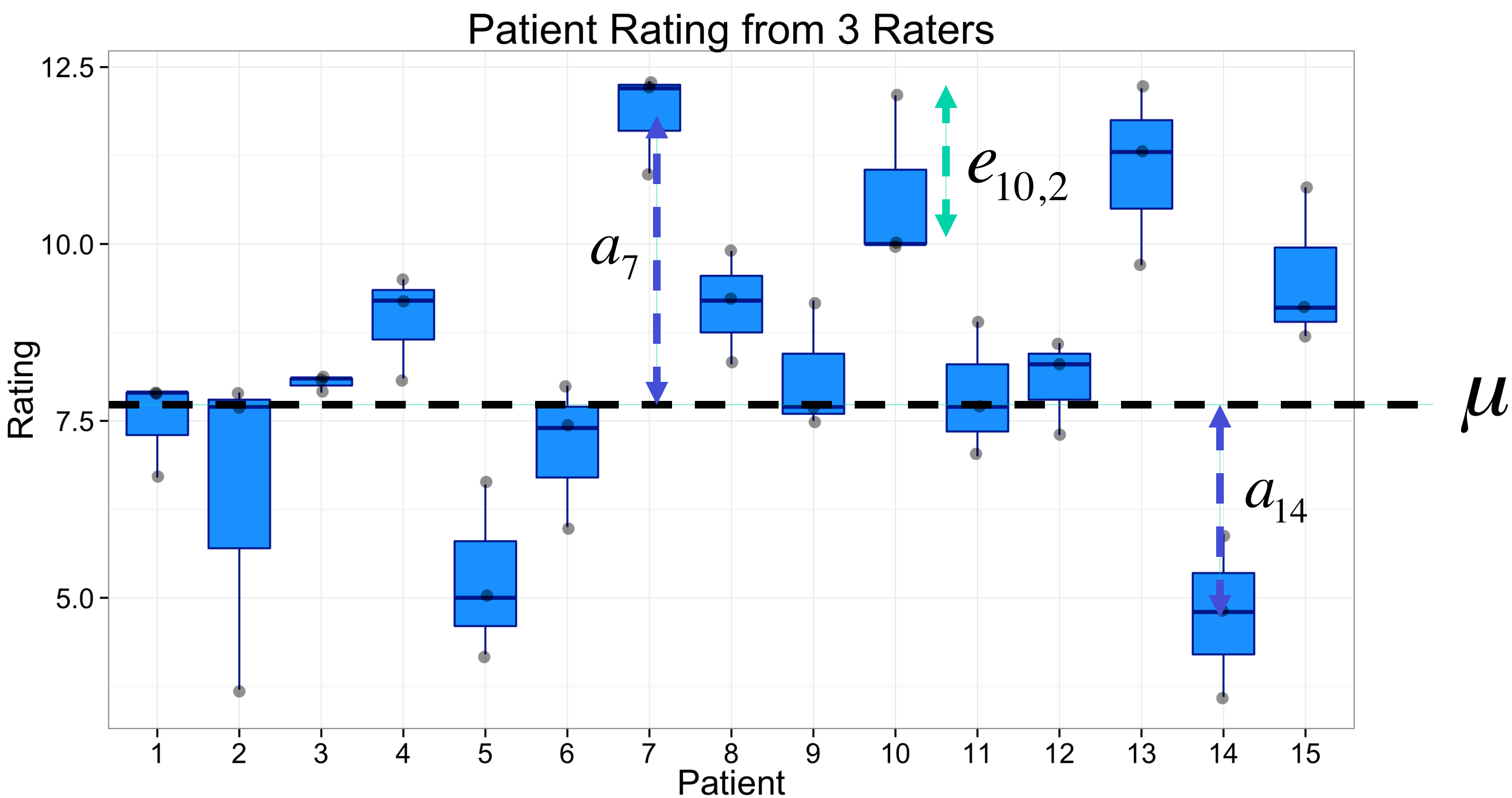
$$Y_{ij} = \mu + a_i + e_{ij}$$



Patient Rating from 3 Raters
12.5
10.0
7.5
5.0
Rating
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Patient
a_7
$e_{10,2}$
μ
a_{14}

$$Y_{ij} = \mu + a_i + e_{ij}$$

$$ICC \propto \frac{\mathrm{var}\ a}{\mathrm{var}\ a + \mathrm{var}\ e}$$



$$\boxed{ICC = \frac{\textit{Group}\ \text{var}}{\textit{Total}\ \text{var}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}. \quad ICC \propto \frac{\text{var}(a)}{\text{var}(a) + \text{var}(e)}}$$

To estimate ICC, we use the Expected Mean Squares (EMS) for the within-group MS and between-group MS:

$E(MS_w) = \sigma_e^2$; $E(MS_b) = \sigma_e^2 + k\sigma_a^2$, ($k$ obs per group).

Thus $\dfrac{E(MS_b) - E(MS_w)}{k} = \sigma_a^2$.

Substituting for σ_a^2 and σ_e^2, and using the observed MS as estimates of $E(MS)$, we get

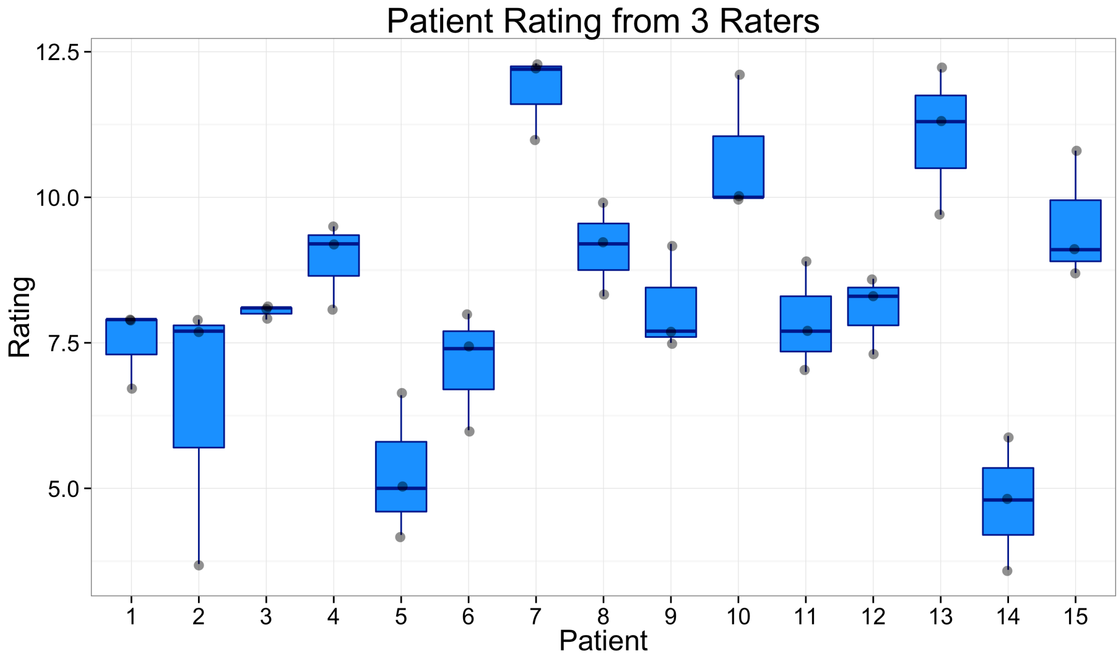
$$ICC = \frac{(MS_b - MS_w)/k}{(MS_b - MS_w)/k + MS_w} = \frac{MS_b - MS_w}{MS_b + (k-1)MS_w}$$

$$= \frac{MS_b / MS_w - 1}{MS_b / MS_w + (k-1)} = \boxed{\frac{F - 1}{F + (k-1)}}$$

- There are many algebraic definitions of *ICC*, all conveying the notion of **within-group similarity relative to between-group similarity**, but differing in the assumed correlation between the group effect, a_i, and the measurement error, e_{ij}.

- There are many algebraic definitions of *ICC*, all conveying the notion of **within-group similarity relative to between-group similarity**, but differing in the assumed correlation between the group effect, a_i, and the measurement error, e_{ij}.
- **‘sintraclass1.r’** shows different formulae for *ICC* for 1- and *k*-item tests. Check out http://www.stanford.edu/class/psych253/tutorials/ICC_from_linearmodels.html for more details.

- There are many algebraic definitions of *ICC*, all conveying the notion of **within-group similarity relative to between-group similarity**, but differing in the assumed correlation between the group effect, a_i, and the measurement error, e_{ij}.
- **'sintraclass1.r'** shows different formulae for *ICC* for 1- and *k*-item tests. Check out http://www.stanford.edu/class/psych253/tutorials/ICC_from_linearmodels.html for more details.
- The formula, *ICC = (F - 1)/(F + k - 1),* shows that, as *F* becomes very large (i.e., *within-group similarity becomes very large relative to between-group similarity*), *ICC* tends to 1. *ICC* = 0, when *F* = 1, i.e., when $MS_w = MS_b$.




```
rs0 = lmer(rating ~ (1|patient), data=d1)
F = 7.63; var_p = 2.34; var_resid = 1.06
icc2 = ICC(d0); icc2[[1]]$ICC[1] #0.69
```

$$ICC = \frac{F-1}{F+(k-1)} = \frac{7.63-1}{7.63+(3-1)} = 0.69$$

$$ICC = \frac{\text{var\_}p}{\text{var\_}p + \text{var\_}resid} = \frac{2.34}{2.34+1.06} = 0.69$$

- ‘Random effects’ models are often more realistic than ‘fixed effects’ models, and they deserve special attention - as do models with both fixed and random effects, known as ‘mixed’ models. (See our Psych 252 notes!)

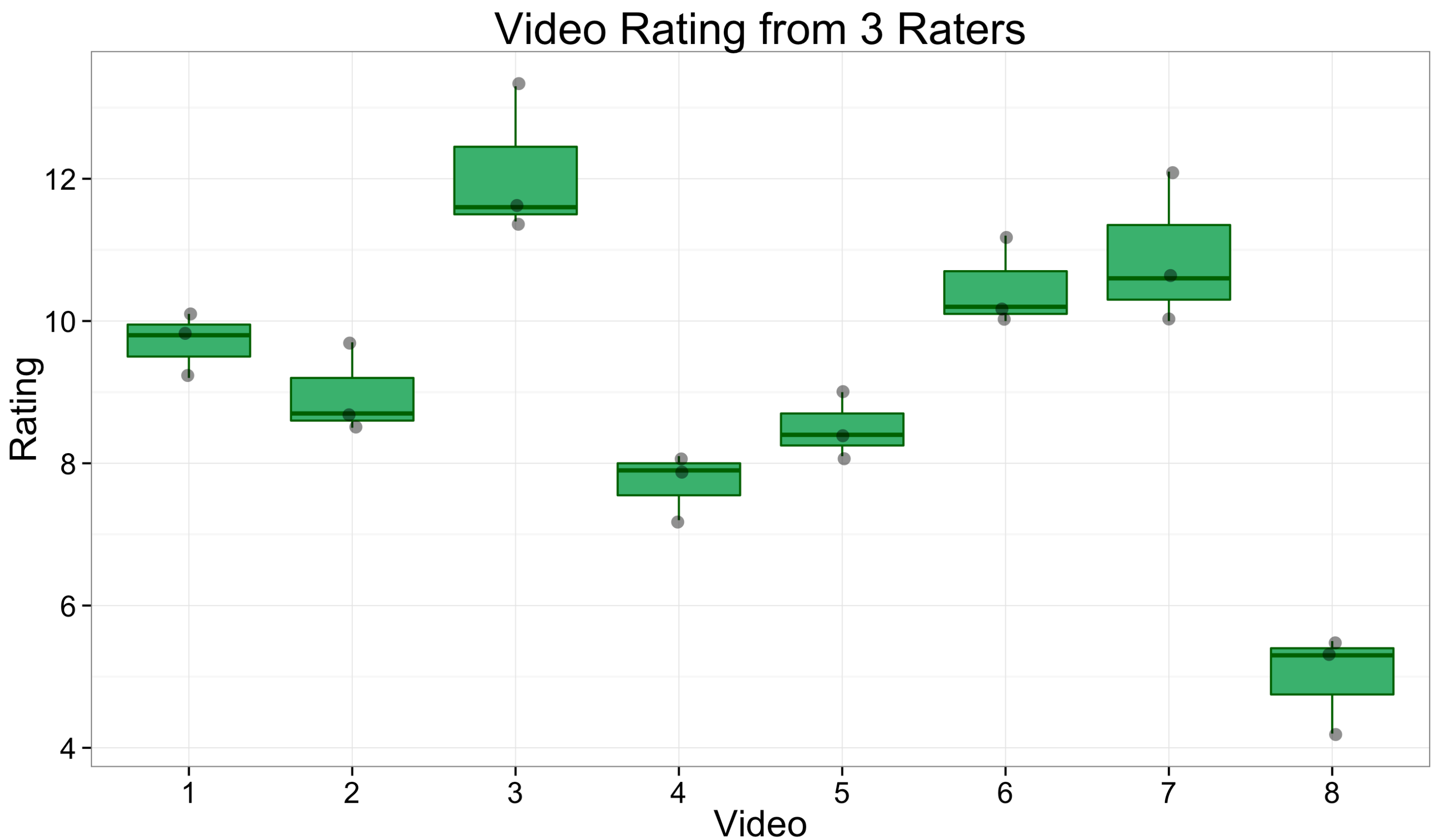
- ‘Random effects’ models are often more realistic than ‘fixed effects’ models, and they deserve special attention - as do models with both fixed and random effects, known as ‘mixed’ models. (See our Psych 252 notes!)
- **Mixed models** can arise in the calculation of reliability for 2-way designs. The **objects being rated are properly seen as a random sample of objects**. The ***k* raters** may or may not be viewed as a **random sample of raters**. E.g., the 9 Justices on the Supreme Court correspond to levels of a fixed-effects factor.

2. 2-way designs. We array the ratings given by the k (= 3) *raters* to p objects, as follows.
Reliability* = *Correlation*, Cronbach's *alpha

	Rater		
P or Object	**1**	**2**	**3**
1 **2**	5 2	2 4	6 4
3 **4**	0 ..	4 ..	2 ..

2. 2-way designs. We array the ratings given by the k (= 3) *raters* to p objects, as follows.

Reliability* = *Correlation*, Cronbach's *alpha



Cronbach’s *alpha* in 2-way tables

- Arrange data into 2-way table, e.g., ‘objects’ as rows, ‘raters’ as columns
- Reliability high if the profile (rise & fall) of scores in a column is the same for all columns; i.e., if the (rank) correl between columns is high; i.e., if the row * column **interaction** is low.
- This is the idea underlying **Cronbach’s *alpha*** (α).

A 2-way ANOVA with n = 1 obs per cell; ***k* raters** (columns) and ***p* objects** (rows).

MS_p = MS for Objects; MS_{res} = MS residual.

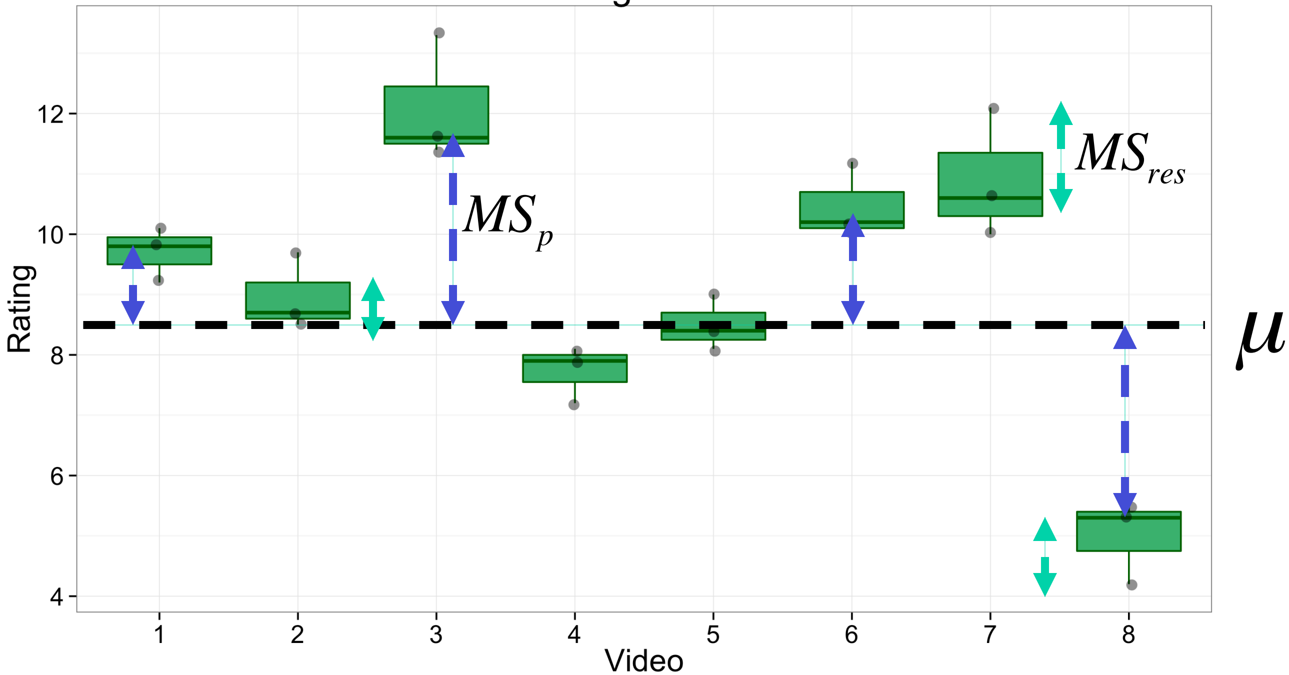
Using the EMS formula for this ANOVA (see the next slide), we can **estimate the variance due to objects, σ_p^2,** as (this is the 'true' variance across Objects's!):

$$\sigma_p^2 = \frac{MS_p - MS_{res}}{k}$$

'Object' is a random effects variable. Should 'rater' be regarded as a random effects variable too?

It depends on the research design.

Video Rating from 3 Raters



$$\sigma_p^2 = \frac{MS_p - MS_{res}}{k} \propto \frac{\quad - \quad}{k}$$

EMS for 2-way design when n = 1

Source	df	E(MS)		
		Fixed	Random	A Fixed, B Random
A(raters)	a-1	$\sigma^2 + b\theta_a$	$\sigma^2 + \sigma^2_{ab} + b\sigma^2_a$	$\sigma^2 + \sigma^2_{ab} + b\theta_a$
B(participants)	b-1	$\sigma^2 + a\theta_b$	$\sigma^2 + \sigma^2_{ab} + a\sigma^2_b$	$\sigma^2 \quad + a\sigma^2_b$
Error	(a-1)(b-1)	$\sigma^2 + \theta_{ab}$	$\sigma^2 + \sigma^2_{ab}$	$\sigma^2 + \sigma^2_{ab}$
Total	ab-1			

In an additive model, put $\theta_{ab} = 0$ or $\sigma^2_{ab} = 0$ (depending on whether the factors are fixed or random). In this case, MSE is the appropriate denominator in the F ratio for testing the 2 main effects.

If A and B are random, MSE is the appropriate denominator in the F ratio for testing **both** main effects, even if $\sigma^2_{ab} > 0$.

p (objects) x *k* (raters) matrix

- In the preceding EMS Table, **let A = 'raters'**, with *a* = *k* levels, let **B = 'objects' / 'participants'**, with *b* = *p* levels, and let us consider the case, "A and B random". Then:

$$MS_p = \sigma^2 + \sigma^2_{rp} + k\sigma^2_p;\ MS_{res} = \sigma^2 + \sigma^2_{rp}.$$

So $\sigma^2_p = \dfrac{MS_p - MS_{res}}{k}$, as stated earlier.

Cronbach defined 'reliability' as the ratio of true score variance to the sum of true score variance and residual variance: $\alpha = \dfrac{\sigma^2_p}{\sigma^2_p + MS_{res}}$.

- Then Cronbach's *alpha* (α) is defined as:

$$\alpha = \frac{\sigma_p^2}{\sigma_p^2 + MS_{res}} = \frac{MS_p - MS_{res}}{MS_p + (k-1)MS_{res}} = \frac{F-1}{F+(k-1)}.$$

- *F* is now defined as MS_p/MS_{res}.
- SPSS gives *alpha*: **Analyze > Scale > Reliability Analysis,** etc.
- If the DV is a **single rater's** score, *reliability* = α.
- If the DV is the **average score for *k* raters**, it will be shown later that *reliability* = $(k\alpha)/[1+(k-1)\alpha]$.

The R function, `ICC()`

```
> str(d0)
 'data.frame':          15 obs. of  3 variables:
 $ rep1: num  9.2 12.1 9.1 5.5 5.7 6.6 6.7 6.4 10.8 4.4 ...
 $ rep2: num  10.7 11.4 11.2 4.4 4.5 4.3 6.9 7.9 11.2 8.6 ...
 $ rep3: num  11 11.4 8 6.3 6.6 4.6 6.7 7 10.5 5 ...
> icc2 = ICC(d0)                                #Ratings from all 3 replicates
> print(icc2)
Call: ICC(x = d0)

Intraclass correlation coefficients
                         type  ICC  F df1 df2       p lower bound upper bound
Single_raters_absolute   ICC1 0.77 11  14  30 3.8e-08        0.55        0.90
Single_random_raters     ICC2 0.76 10  14  28 1.4e-07        0.54        0.91
Single_fixed_raters      ICC3 0.75 10  14  28 1.4e-07        0.52        0.90
Average_raters_absolute ICC1k 0.91 11  14  30 3.8e-08        0.78        0.97
Average_random_raters   ICC2k 0.91 10  14  28 1.4e-07        0.78        0.97
Average_fixed_raters    ICC3k 0.90 10  14  28 1.4e-07        0.77        0.96

 Number of subjects = 15     Number of Judges =  3
```

Single ratings

Avg of k raters (less variance, so higher ICCs)

Absolute vs. Fixed vs. Random

- **Absolute:** Different rater rates each object (randomly sample raters); *1-way random effects ANOVA model*
- **Random**: Each rater rates each object, and rater treated as random effect; can generalize to greater population of judges!; *2-way random effects model*
- **Fixed:** Each rater rates each object, and rater treated as fixed effect; no generalization; *2-way mixed effects model*

The R function, `kappam.light()`

- The package, **irr**, seems to be very flexible (moreso than **psych**), containing many indices of reliability. However, it does not accept an agreement matrix as input, only raw data. For agreement matrices, use `cohen.kappa()` in **psych**. (See HW-1 for details.)
- Compare results of `kappam.light(),` when data are quantitative, with ***alpha***. Do they agree?

4. Pivot to theory-based approaches to reliability

- For many (4, at least) data ‘formats’, we know how to compute reliability.
- Theory-based approaches rely on a very general model, **Test Theory**, in which a **‘test’** consists of many **‘items’**, and we wish to express the reliability of the ‘test’ as a function of that of ‘items’
- Define *reliability* as *internal consistency* of a ‘test’.

Generality of ‘item/test’ model

- This model of ‘tests’ as consisting of ‘items’ is very general:
  - Averaging across trials of BOLD activity to get more reliable signal of brain activity
  - Measuring a construct by 2 or more methods
  - Using *k* semantic differential scales to measure ‘thought valence’ or ‘attitude’
- Thus interest in ‘reliability’ is, or ought to be, widespread

Summary of Theory-based results

Method

	Rater		
P or Object	**1**	**2**	**3**
1 **2**	5 2	2 4	6 4
3 **4**	0 ..	4 ..	2 ..

Summary of Theory-based results

Method

	Rater		
P or Object	**1**	**2**	**3**
1 **2**	5 2	2 4	6 4
3 **4**	0 ..	4 ..	2 ..

Test

	Item		
Person	**1**	**2**	**3**
1 **2**	5 2	2 4	6 4
3 **4**	0 ..	4 ..	2 ..

Summary of Theory-based results

- ‘Test’ has *k (*or *n)* ‘items’ for measuring a construct on ‘persons’. Or, analogously, replace ‘test’ by ‘method’, ‘item’ by ‘rater’, and ‘person’ by ‘object’.
- **Qu**: Is **test** reliable?
- Reword question as: “Is test internally consistent, i.e., do the **items** “hang together”, i.e., is the correl, ρ, between items across persons ‘not low’?”
- **Ans to Qu: It depends on $\boldsymbol{\rho}$ and *k***.

Summary (cont'd)

- Reliability of a test *increases* **with the length**, k, and the **internal consistency**, ρ, of the test (Spearman-Brown formula)
- Reliability is *reduced* when we **reduce the range of *X***
- The **estimated correlation**, r_{AY}, between some 'interesting' variable, Y, and the *latent* construct, A, that is measured by X, *increases* as the **reliability of *X* increases**.

Appendix

- R script, ‘sintraclass1.r’, showing different versions of ICC, each justified by its own ANOVA model

	rep1	rep2	rep3
P1	8.0	6.2	8.4
P2	7.4	9.8	8.5
P3	10.4	7.5	9.1
P4	4.6	3.7	5.6
P5	9.5	9.7	9.9
P6	9.2	8.6	8.6
P7	5.3	4.7	5.5
P8	6.6	7.9	6.9
P9	4.2	6.9	3.6
P10	5.4	6.4	6.9
P11	6.9	6.9	10.2
P12	10.6	10.6	11.6
P13	11.5	11.6	11.3
P14	6.3	8.2	6.8
P15	7.5	9.7	9.0

- (Data file in ‘short’ form, but note that “rep1” may **not** be the same rater in the different rows! The ‘long’ form needed for ANOVA is on next slide, with 45 rows and each column being a variable or factor)

	rating	patient	replic (== rater)
1	8.0	1	1
2	7.4	2	1
3	10.4	3	1
..	..	..	..
16	6.2	1	2
17	9.8	2	2
18	7.5	3	2
..	..	..	..
9	8.4	1	3
10	8.5	2	3
11	9.1	3	3
..	..	..	..
43	11.3	13	3
44	6.8	14	3
45	9.0	15	3

R script, ‘sintraclass1.r’

```
#Script to generate the patient-as-random-effect data
   in a 15x3 2-way design with n=1.  Then check that
   various formulae for ICC give same results.

library(lme4)   #Need to install 'lme4' package for mixed models ANOVA
library(psych)   #contains ICC()

e0 = matrix(rnorm(45), ncol=3) #Errors for 15x3 matrix
p0 = matrix(2*rnorm(15), ncol=3, nrow=15)  #Patient random effects
d0 = round(data.frame(8 + p0 + e0), 1)  #Obs scores
rownames(d0) = paste('P',1:15,sep='')
colnames(d0) = paste('rep',1:3,sep='')

#Data in ‘long’ form for random effects model
d00 = as.matrix(d0)
d1 = data.frame(cbind(c(d00), rep(1:15,3), rep(1:3, each=15)))
colnames(d1) = c("rating","patient","replicate")
```

```
#Reverse order (D1,D2) pairs and find cor. Use ICC() to check.
x1 = c(d0[,1], d0[,2]); y1 = c(d0[,2], d0[,1])
icc0 = round(cor(x1, y1), 2)
cat('ICC12 as correl across 2n pairs = ', icc0)

d01 = d0[,1:2]   #D1, D2 ratings
icc1 = ICC(d01)  #ICC() returns a list; icc1[[1]] contains ICC
cat('ICC12, judges fixed, = ', icc1[[1]]$ICC[3])
   #icc1[[1]]$ICC[3] should = icc0 (correlation across pairs)
cat('ICC12, judges random, = ', icc1[[1]]$ICC[2])
   #icc1[[1]]$ICC[3] = icc1[[1]]$ICC[2]??
   #Also calculate the average of ICC12, ICC13 & ICC23; compare with
   icc2 below

icc2 = ICC(d0)    #Reliability for ratings from all 3 doctors
cat("Whole sample ICC = ", icc2$ICC[2])
```

Results for a similar data set with 3 raters:

- ICC12 as correl across 2n pairs = 0.77
- ICC12, judges fixed, = 0.77
- ICC12, judges random, = 0.78
- Whole sample ICC = 0.7
- Average ICC = 0.7 (ICC13=.6, ICC23=.73)

```
cat('Compute ICC from ANOVA table')
cat('ICC = (F-1)/(F+k-1)')

rs0 = lm(rating ~ patient, data=d1)    #This fixed-effects
   model is not exactly right, but it gives the right F
print(anova(rs0))
```

```
Results:
Compute ICC from ANOVA tables: ICC = (F-1)/(F+k-1)
Analysis of Variance Table

Response: rating
          Df Sum Sq Mean Sq F value    Pr(>F)
patient   14 181.77  12.984  8.0945 9.527e-07 ***
Residuals 30  48.12   1.604
```

```
(F-1)/(F+k-1) = (8.09-1)/(8.09+3-1) = 0.70,
which agrees with the 'whole sample' ICC.
```

```
cat('Compute ICC from ANOVA table')
cat('ICC = var(p)/[var(p)+var(resid)]')

rs1 = lmer(rating ~ (1 | patient), data=d1)   #This
   random-effects model is the right model
print(summary(rs1))
```

```
Results:
Linear mixed model fit by REML
Formula: rating ~ (1 | patient)
Random effects:
 Groups   Name          Variance Std.Dev.
patient  (Intercept)   3.7932   1.9476
Residual               1.6040   1.2665
Number of obs: 45, groups: patient, 15

ICC = var(p)/[var(p)+var(resid)] = 3.79/(3.79+1.60) =
   0.70, as before
```

- Given in the `ICC()` output are the reliability of a test consisting of 1 item, and that of a test consisting of *k* items.
- The relation between these 2 indices is given by the **Spearman-Brown formula**, which will be derived later on.
- Most often, the index of choice is Cronbach's *alpha* for a test with *k* items. This is **ICC3K** in the output from `ICC()`.