University of Colorado
Boulder

Human-Robot Interaction

Analysis of Variance

Professor **Dan Szafir**

Computer Science & ATLAS Institute
University of Colorado Boulder

Review

Paired or Matched Observations

Variable	Test
Nominal	McNemar’s Test
Ordinal (ordered categories)	Wilcoxon
Quantitative (discrete or non-normal)	Wilcoxon
Quantitative (normal)	Paired t-test

Independent Observations

Output Variable

Input Variable

	Nominal	Categorical (>2 categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Nominal	X^2 or Fisher's	X^2	X^2 trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank (a)	Student's t-test
Categorical (>2 categories)	X^2	X^2	Kruskal-Wallis (b)	Kruskal-Wallis (b)	Kruskal-Wallis (b)	Analysis of variance (c)
Ordinal	X^2 trend or Mann-Whitney	(e)	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression (d)
Quantitative Discrete	Logistic regression	(e)	(e)	Spearman rank	Spearman rank	Spearman rank or linear regression (d)
Quantitative Non-Normal	Logistic regression	(e)	(e)	(e)	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
Quantitative Normal	Logistic regression	(e)	(e)	(e)	Linear regression (d)	Pearson and linear regression

Why not use T-tests all the time?

Multiple Comparisons

T-test works well for comparing two populations

- Example: single factor, 2 levels

What do you do when you have multiple categories?

Multiple Comparisons

T-test works well for comparing two populations

- Example: single factor, 2 levels

What do you do when you have multiple categories?

- Idea: conduct multiple pair-wise t-tests
- Problem: Type 1 error (rejecting H_0 when it is true)
- E.g.,: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, Confidence interval: 0.95
- 6 pairwise tests – $(0.95)^6 = 0.74$
  - Equivalent to rejecting H_0 when $p<0.26$ instead of when $p<0.05$

Errors in Hypothesis Testing

Types of Errors

Type 1

Reject null hypothesis when it is true

Type 2

Accept the null hypothesis when it is false

Type 3

Correctly rejecting the null hypothesis for the wrong reason

Hypothesis Testing

	Null Hypothesis (H0) is true	Alternative Hypothesis (H1) is true
Fail to Reject Null Hypothesis	*Right decision*	**Type II Error** Wrong decision False Negative
Reject Null Hypothesis	**Type I Error** Wrong decision False Positive	*Right decision*

Output Variable

Input Variable

	Nominal	Categorical (>2 categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Nominal	X^2 or Fisher's	X^2	X^2 trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank (a)	Student's t-test
Categorical (>2 categories)	X^2	X^2	Kruskal-Wallis (b)	Kruskal-Wallis (b)	Kruskal-Wallis (b)	**Analysis of variance (c)**
Ordinal	X^2 trend or Mann-Whitney	(e)	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression (d)
Quantitative Discrete	Logistic regression	(e)	(e)	Spearman rank	Spearman rank	Spearman rank or linear regression (d)
Quantitative Non-Normal	Logistic regression	(e)	(e)	(e)	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
Quantitative Normal	Logistic regression	(e)	(e)	(e)	Linear regression (d)	Pearson and linear regression

Analysis of Variance (ANOVA)

Types of ANOVA

- Different procedures
  - One-way ANOVA – single factor, two or more levels
    - Single factor, two level ANOVA is equivalent to a t-test
  - Two-way ANOVA – two factors, two or more levels for each
  - Multi-way ANOVA – multiple factors
- Different ways of modeling
  - Fixed effects, random effects, mixed effect
  - Independent or paired data

High level picture of ANOVA

Calculate the variance within groups

Calculate how the within-group variance would translate into differences between groups

Take into account how many subjects there are in each group

If the observed differences are a lot bigger than what you'd expect by chance, you have statistical significance

More formally

ANOVA relies on the Fisher (F) ratio

F = variance between groups / variance within sample group

$$\sum n_i \left(M_i - \sum \left(\frac{M_i}{k} \right) \right)^2$$ = Sum of squares for treatment

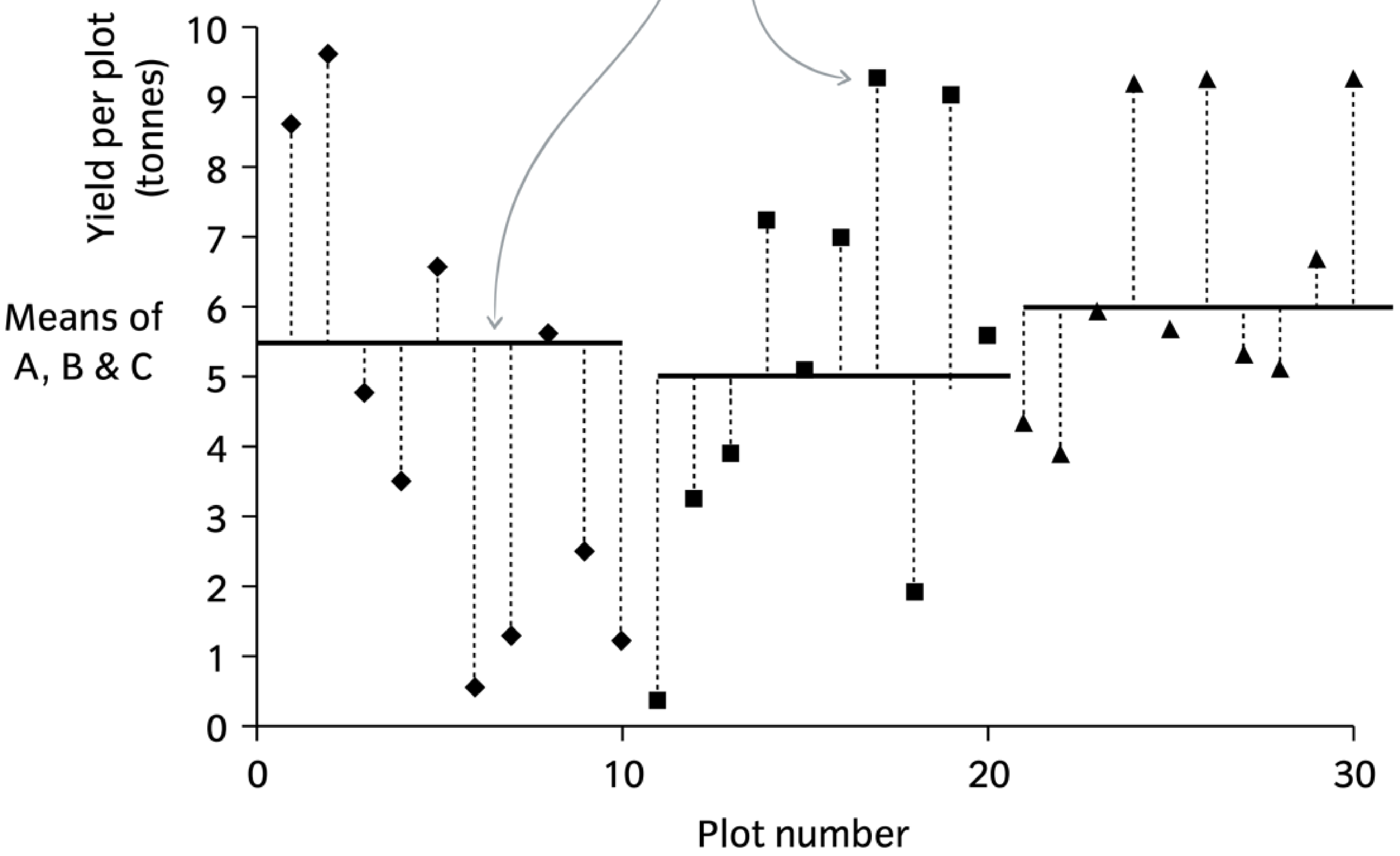
$$\boldsymbol{F} = \frac{explained\ variance}{unexplained\ variance} = \frac{SST/(k-1)}{SSE/(n-k)}$$

populations

total sample size

$$\sum \sum (X_{it} - M_i)^2$$ = Sum of squares for error

$$F = \frac{\sum n_i(M_i - \sum(M_i/k))^2/(k-1)}{\sum\sum(X_{it} - M_i)^2\ /(n-k)}$$



Types of ANOVA

- Different procedures
  - **One-way ANOVA – single factor, two or more levels**
    - Single factor, two level ANOVA is equivalent to a t-test
  - Two-way ANOVA – two factors, two or more levels for each
  - Multi-way ANOVA – multiple factors
- Different ways of modeling
  - Fixed effects, random effects, mixed effect
  - Independent or paired data

One-way ANOVA

Between-Participants One-Way ANOVA in R

Between-subjects comparisons

Example: Participants are given different doses of a drug. We are interested in how these dosages affect alertness.

```
> data.ow.b=read.table("http://personality-project.org/r/datasets/R.appendix1.data",header=TRUE)
> fit.ow.b = aov(Alertness~Dosage, data=data.ow.b)
> summary(fit.ow.b)
            Df  Sum Sq  Mean Sq  F value   Pr(>F)
Dosage       2   426.2   213.12    8.789  0.00298 **
Residuals   15   363.8    24.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit <- aov(y~B,data)
```

Interpreting an ANOVA Table

One-way Analysis of Variance

Source	DF	SS	MS	F	P
Factor	m-1	SS(Between)	MSB	MSB/MSE	
Error	n-m	SS(Error)	MSE		
Total	n-1	SS(Total)			

From F-distribution with m-1 numerator and n-m denominator d.f.

MSB = SS(Between)/(m-1)
MSE = SS(Error)/(n-m)

n-1 = (m-1) + (n-m)

SS(Total) = SS(Between) + SS(Error)

Interpreting Results of Between-Participants One-Way ANOVA in R

```
              Df  Sum Sq  Mean Sq   F value  Pr(>F)
Dosage         2   426.2   213.12    8.789   0.00298 **
Residuals     15   363.8    24.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F(2,15) = 8.79$, $p<0.01$ – Reject H_0

Degrees of Freedom

F(**2**,**15**) = 8.79, $p<0.01$ – Reject H_0

DF_1 = k – 1 = # groups – 1

DF_2 = n – k = total population size – # of groups

Interpreting Results

Small p-value:

- It is unlikely that the differences you observed are due to random sampling
- You can reject the idea that all the populations are samples drawn from the same underlying distribution
  - But where is the difference?

Interpreting Results

Small p-value:

- It is unlikely that the differences you observed are due to random sampling
- You can reject the idea that all the populations are samples drawn from the same underlying distribution
  - But where is the difference?
- That doesn't mean that every mean differs from every other mean! Just at least one differs from the rest
  - Conduct **post-hoc** tests to identify where the differences are
  - Also need to worry about Type I and Type II errors in your post-hoc tests!

Post-Hoc Comparisons

Different methods exist that balance trade-offs in risk of Type 1 and Type 2 errors

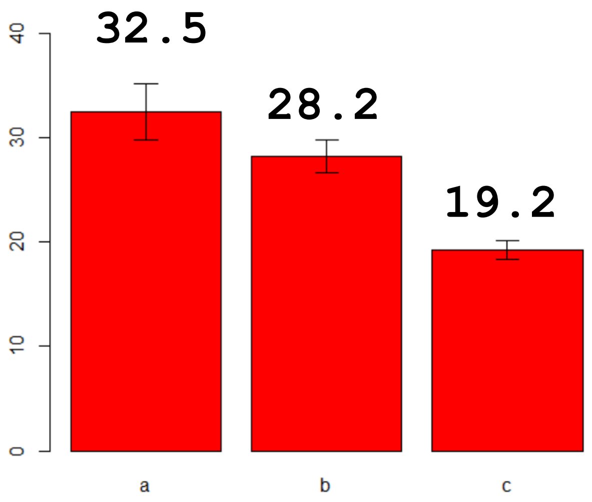
- Tukey’s Honest Significant Difference test (Tuykey’s HSD)
- Bonferroni corrections
- Scheffé's method
- Dunnett’s test
- …

Post-Hoc Comparisons in R
Step 1: Look at the data

```
> print(model.tables(fit.ow.b,"means"),digits=3)
Tables of means
Grand mean
27.66667

 Dosage
          a     b     c
       32.5  28.2  19.2
rep     6.0   8.0   4.0
```

Post-Hoc Comparisons in R
Step 1: Look at the data

```
> library(sciplot)

> bargraph.CI(data.ow.b$Dosage,
data.ow.b$Alertness, ylim=c(0,45),col="Red")
```

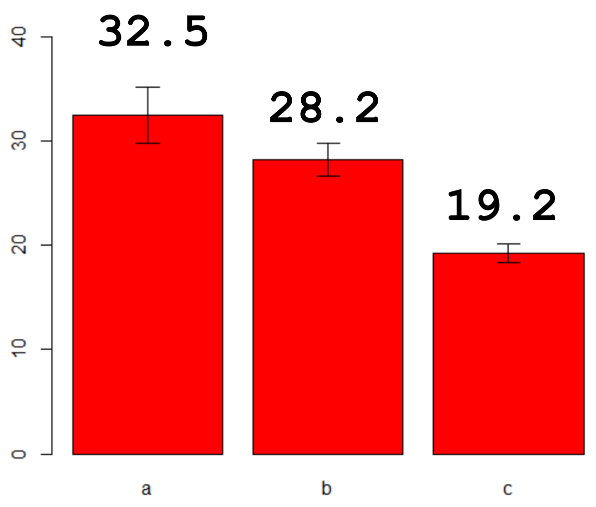


Post-Hoc Comparisons in R
Step 2: Perform desired test

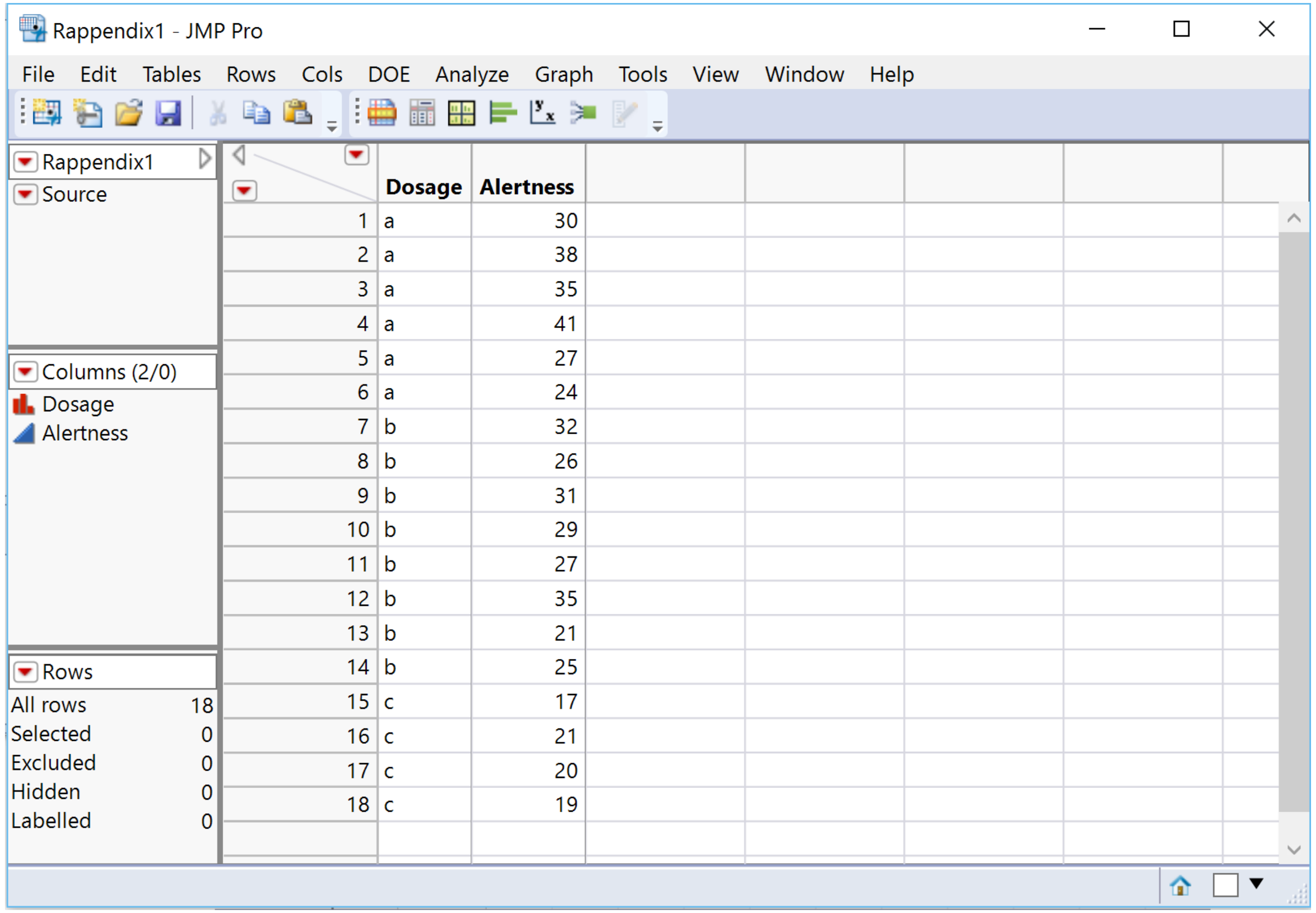
Tukey's HSD

Essentially a series of all possible pairwise t-tests that adjusts p-value to correct for experiment-wise error rate

```
> TukeyHSD(fit.ow.b)

  Tukey multiple comparisons of means

    95% family-wise confidence level

Fit: aov(formula = Alertness ~ Dosage, data = data.ow.b)

$Dosage

      diff       lwr       upr     p adj

b-a  -4.25 -11.15796  2.657961 0.2768132

c-a -13.25 -21.50659 -4.993408 0.0022342

c-b  -9.00 -16.83289 -1.167109 0.0237003
```

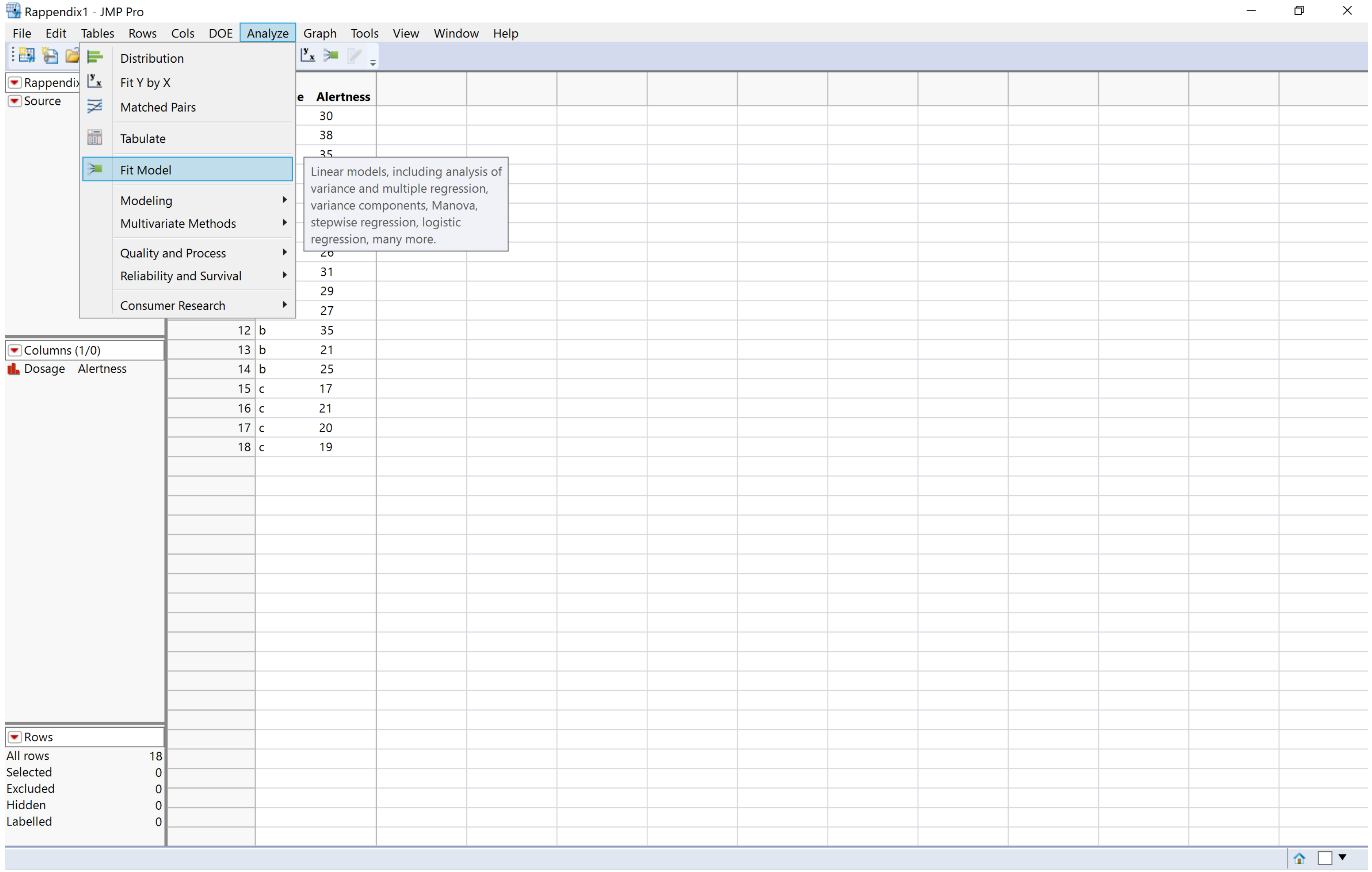


Between-Participants One-Way ANOVA in JMP

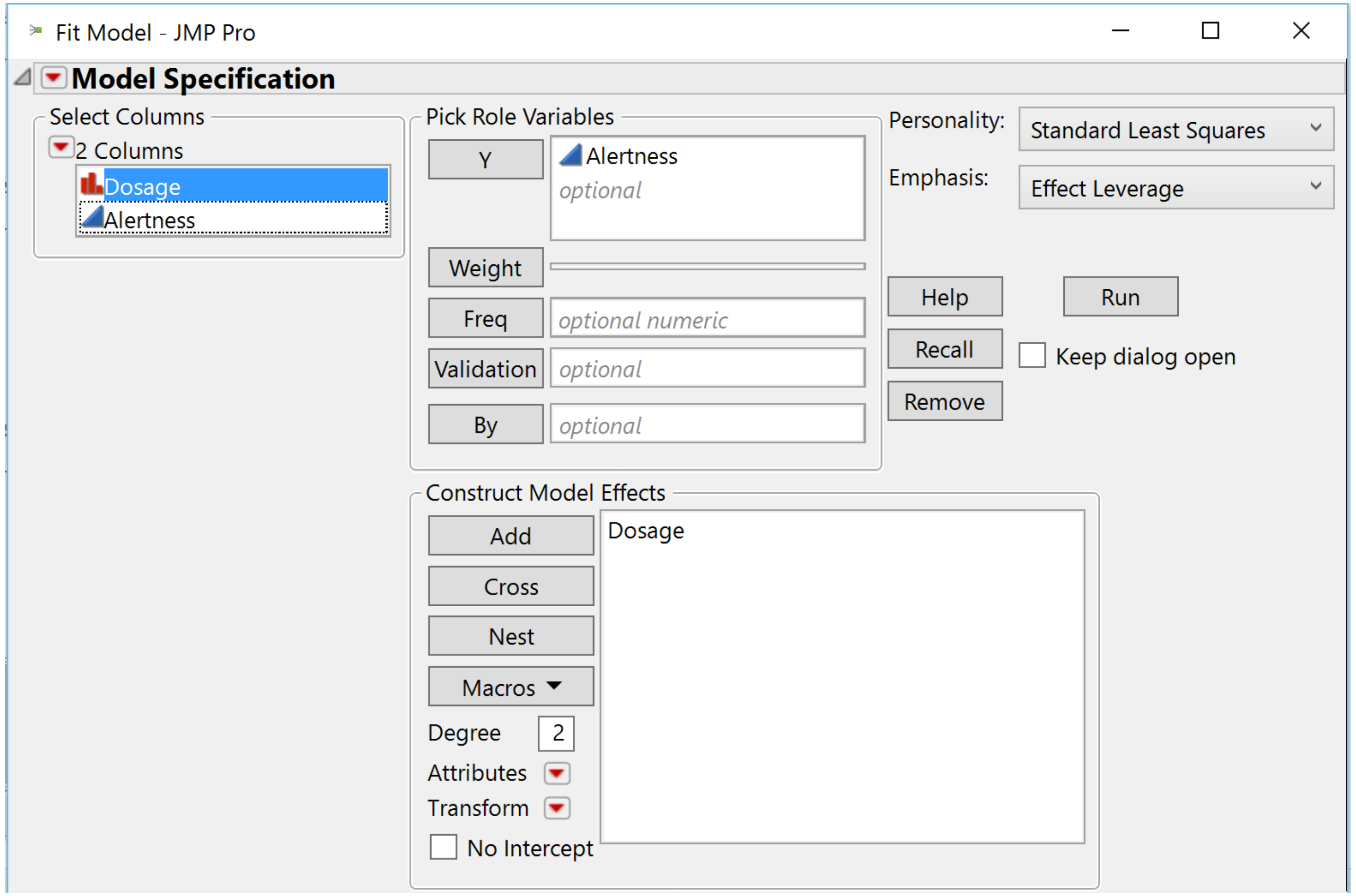


	Dosage	Alertness
1	a	30
2	a	38
3	a	35
4	a	41
5	a	27
6	a	24
7	b	32
8	b	26
9	b	31
10	b	29
11	b	27
12	b	35
13	b	21
14	b	25
15	c	17
16	c	21
17	c	20
18	c	19

Between-Participants One-Way ANOVA in JMP



Between-Participants One-Way ANOVA in JMP

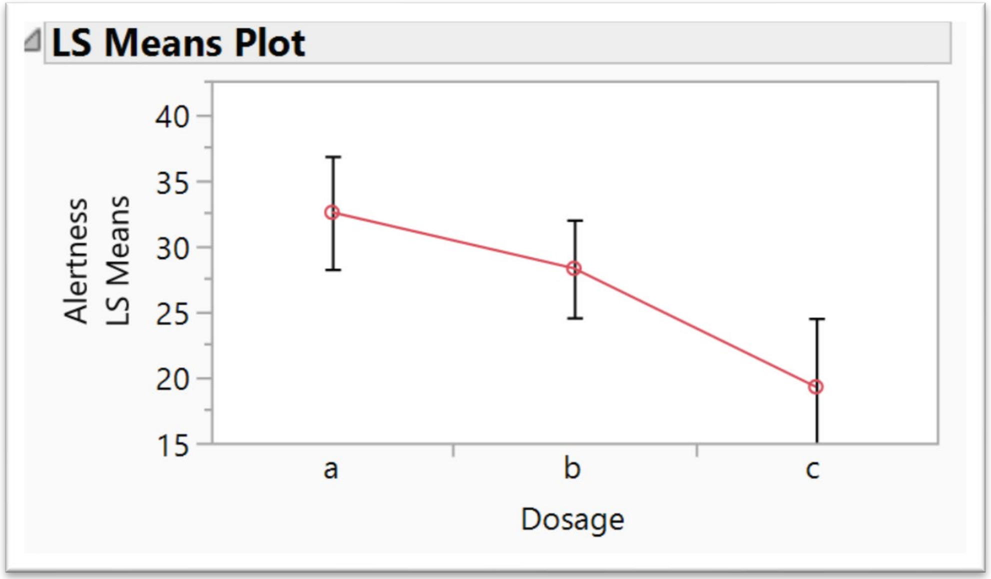


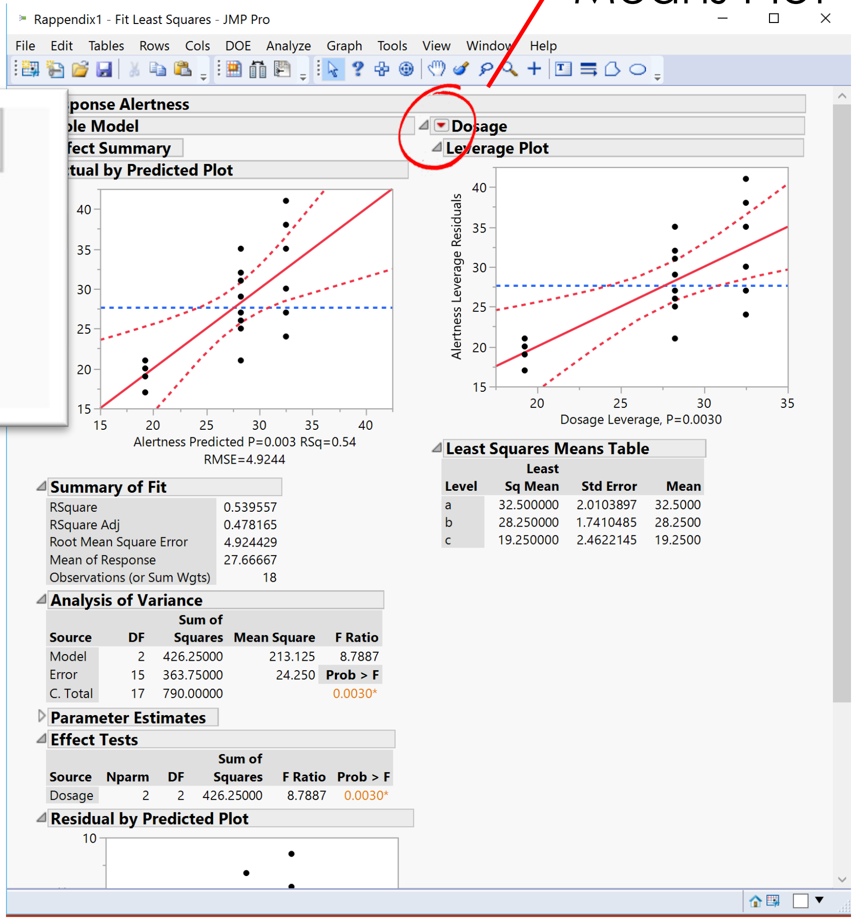
Between-Participants One-Way ANOVA in JMP

To get LS Means Plot

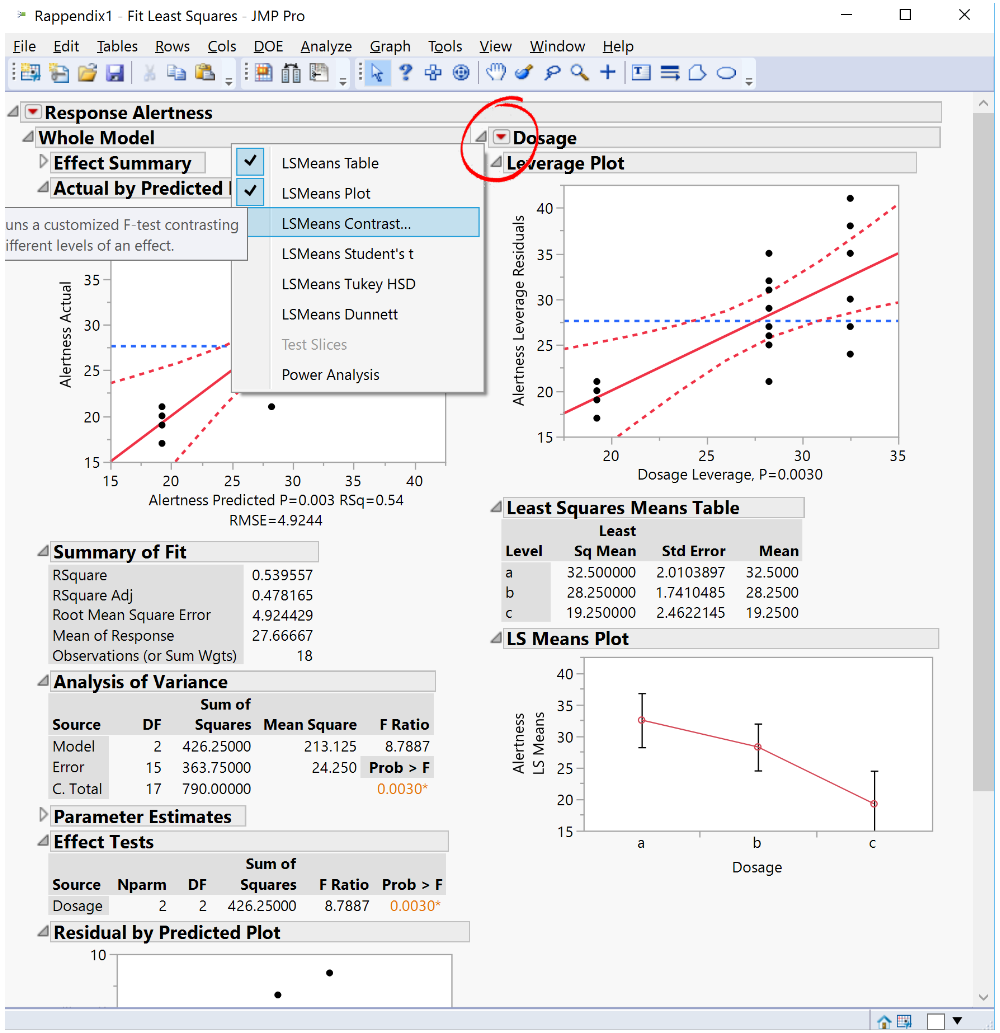
Effect Tests

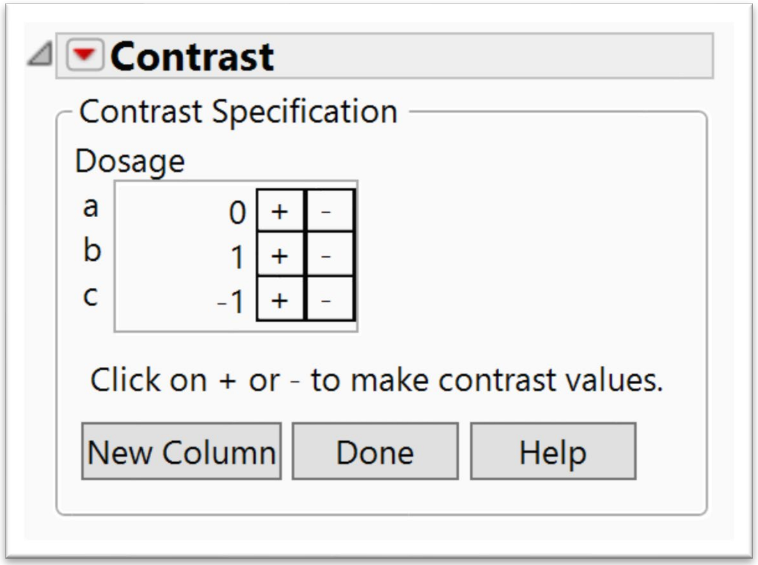
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Dosage	2	2	426.25000	8.7887	0.0030*

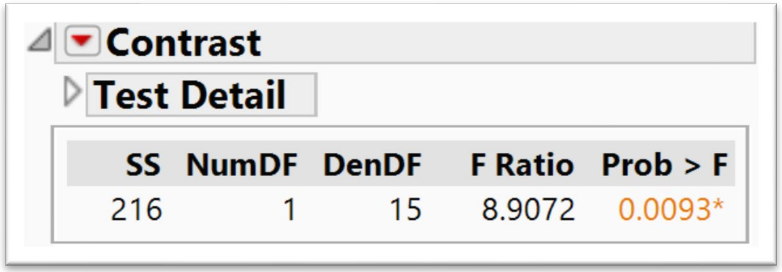




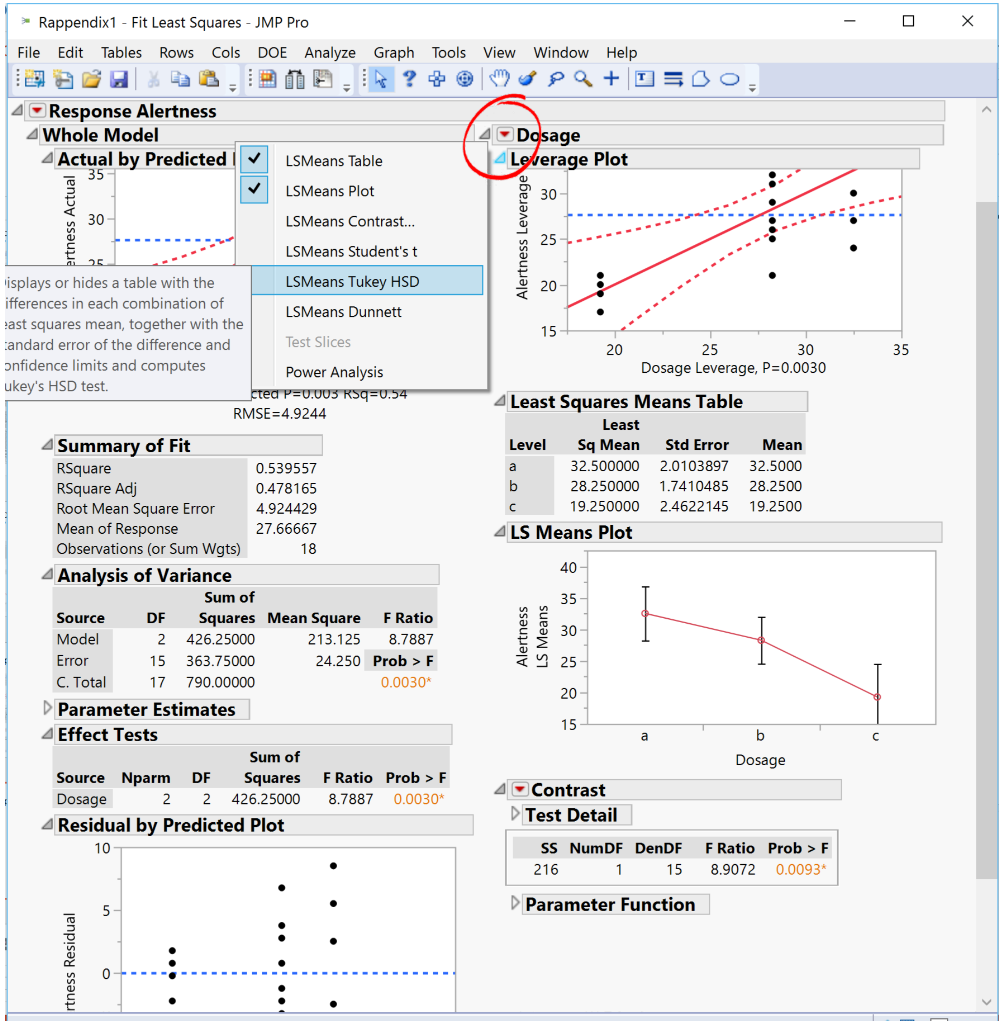
Post-Hoc Comparisons in JMP







Post-Hoc Comparisons in JMP



Post-Hoc Comparisons in JMP

To get Ordered Difference Report

LSMeans Differences Tukey HSD

α= 0.050 Q= 2.59747

LSMean[j]

LSMean[i]	Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	a	b	c
	a	0	4.25	13.25
		0	2.6595	3.17871
		0	-2.658	4.99341
		0	11.158	21.5066
	b	-4.25	0	9
		2.6595	0	3.01558
		-11.158	0	1.16711
		2.65796	0	16.8329
	c	-13.25	-9	0
		3.17871	3.01558	0
		-21.507	-16.833	0
		-4.9934	-1.1671	0

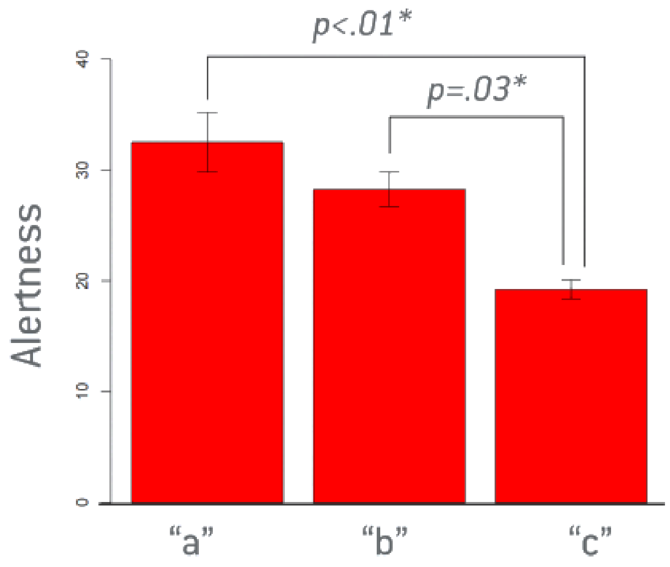
Level			Least Sq Mean
a	A		32.500000
b	A		28.250000
c		B	19.250000

Levels not connected by same letter are significantly different.

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
a	c	13.25000	3.178705	4.99341	21.50659	0.0022*
b	c	9.00000	3.015585	1.16711	16.83289	0.0237*
a	b	4.25000	2.659496	-2.65796	11.15796	0.2768

Reporting Results

"We conducted a one-way analysis of variance (ANOVA) to test whether the different drug dosages result in different levels of alertness and found a significant effect of dosage on alertness $F(2,15) = 8.79$, $p < 0.01$. Post-hoc comparisons using Tukey's HSD test revealed that dosage "c" resulted in significantly different levels of alertness with dosage "b" ($p = .03$) and with dosage "a" ($p < .01$)."



Repeated Measures ANOVA

Within-Participants One-Way ANOVA

Example: participants asked to memorize three types of words with different levels of valence: *positive*, *negative*, and *neutral*. Test word recall by valence.

```
> data.ow.w=read.table("http://personality-project.org/r/datasets/R.appendix3.data",header=TRUE)

> data.ow.w
```

	Observation	Subject	Valence	Recall
1	1	Jim	Neg	32
2	2	Jim	Neu	15
3	3	Jim	Pos	45
4	4	Victor	Neg	30
5	5	Victor	Neu	13
6	6	Victor	Pos	40
7	7	Faye	Neg	26
8	8	Faye	Neu	12
9	9	Faye	Pos	42
10	10	Ron	Neg	22
11	11	Ron	Neu	10
12	12	Ron	Pos	38
13	13	Jason	Neg	29
14	14	Jason	Neu	8
15	15	Jason	Pos	35

Within-Participants One-Way ANOVA in R

```
> fit.ow.w = aov(Recall~Valence+Error(Subject/Valence), data=data.ow.w)

> summary(fit.ow.w)

Error: Subject

          Df Sum Sq Mean Sq F value Pr(>F)

Residuals  4  105.1   26.27

Error: Subject:Valence

          Df Sum Sq Mean Sq F value   Pr(>F)

Valence    2 2029.7  1014.9   189.1 1.84e-07 ***

Residuals  8   42.9     5.4

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit <- aov(y~W+Error(Subject/W),data)
```

Interpreting Results

```
Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4  105.1   26.27
Error: Subject:Valence
          Df Sum Sq Mean Sq F value   Pr(>F)
Valence    2 2029.7  1014.9   189.1 1.84e-07 ***
Residuals  8   42.9     5.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

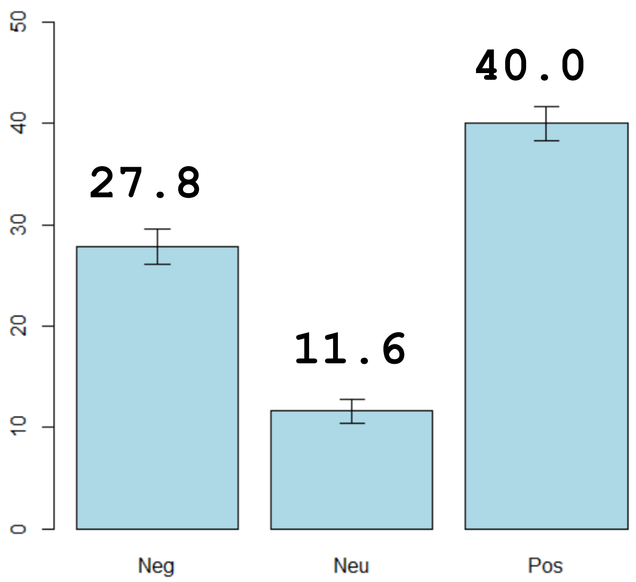
$F(2,8) = 189.1$, $p<.001$ – Reject H_0

Within-Participants One-Way ANOVA Post-Hoc Comparisons in R

```
> print(model.tables(fit.ow.w,"means"),digits=3)

Tables of means

Grand mean


26.46667


 Valence

Valence

 Neg  Neu  Pos

27.8 11.6 40.0
```

Within-Participants One-Way ANOVA Post-Hoc Comparisons in R

```
> library(sciplot)

> bargraph.CI(data.ow.w$Valence,
data.ow.w$Recall, ylim=c(0,50),col="Light Blue")
```



Within-Participants One-Way ANOVA Post-Hoc Comparisons in R

More complicated in R due to table setup

- Need to use the linear model (with libraries “nlme” and “multcomp”)
- I suggest it is easier to just use JMP

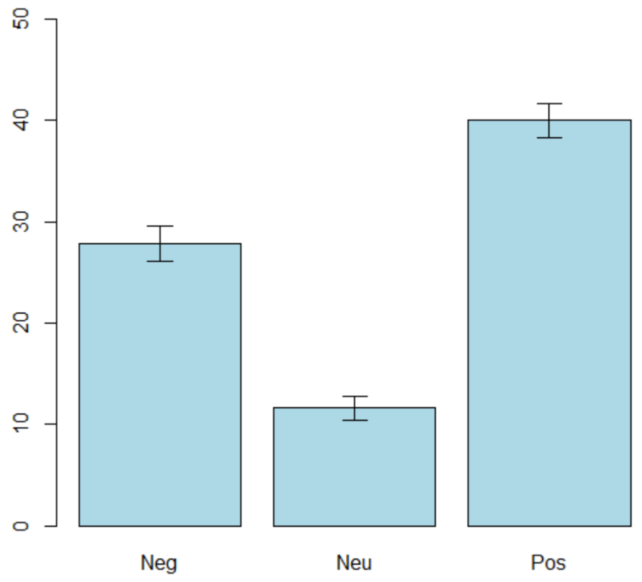
Using the linear model

Within-Participants One-Way ANOVA **Main effect** in R

```
> library(nlme)
> library(multcomp)
> model = aov(Recall~Valence+Error(Subject/Valence), data=data.ow.w)
> summary(model)
Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4  105.1   26.27
Error: Subject:Valence
          Df Sum Sq Mean Sq F value   Pr(>F)
Valence    2 2029.7  1014.9   189.1 1.84e-07 ***
Residuals  8   42.9     5.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Within-Participants One-Way ANOVA
Post-Hoc Comparisons in R

```
> model2 = lme(Recall~Valence, random = ~1|Subject/Valence, data=data.ow.w)
> anova(model2)
            numDF denDF  F-value p-value
(Intercept)     1     8 400.0243  <.0001
Valence         2     8 189.1052  <.0001
> model3 <- glht(model2, linfct=mcp(Valence="Tukey"))
> summary(model3)
Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lme.formula(fixed = Recall ~ Valence, data = data.ow.w, random = ~1 |
    Subject/Valence)
Linear Hypotheses:
               Estimate Std. Error z value Pr(>|z|)
Neu - Neg == 0  -16.200      1.465 -11.057   <1e-10 ***
Pos - Neg == 0   12.200      1.465   8.327   <1e-10 ***
Pos - Neu == 0   28.400      1.465  19.384   <1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```



Within-Participants One-Way ANOVA

Main effect:

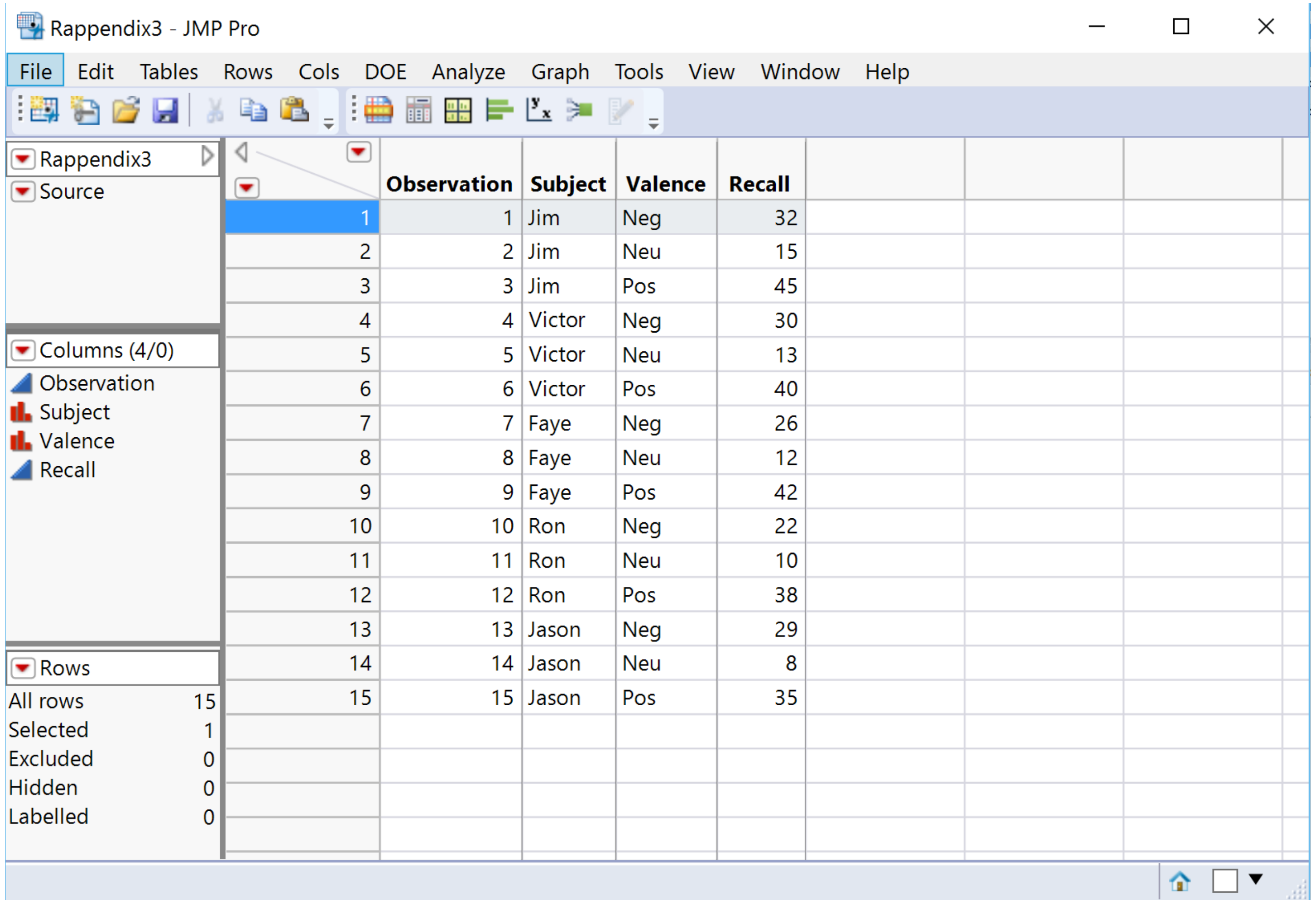
```
fit <- aov(y~W+Error(Subject/W),data)
```

Post-hoc comparisons:

```
linearM <- lme(y~W, random = ~1 | Subject/W),data)
```

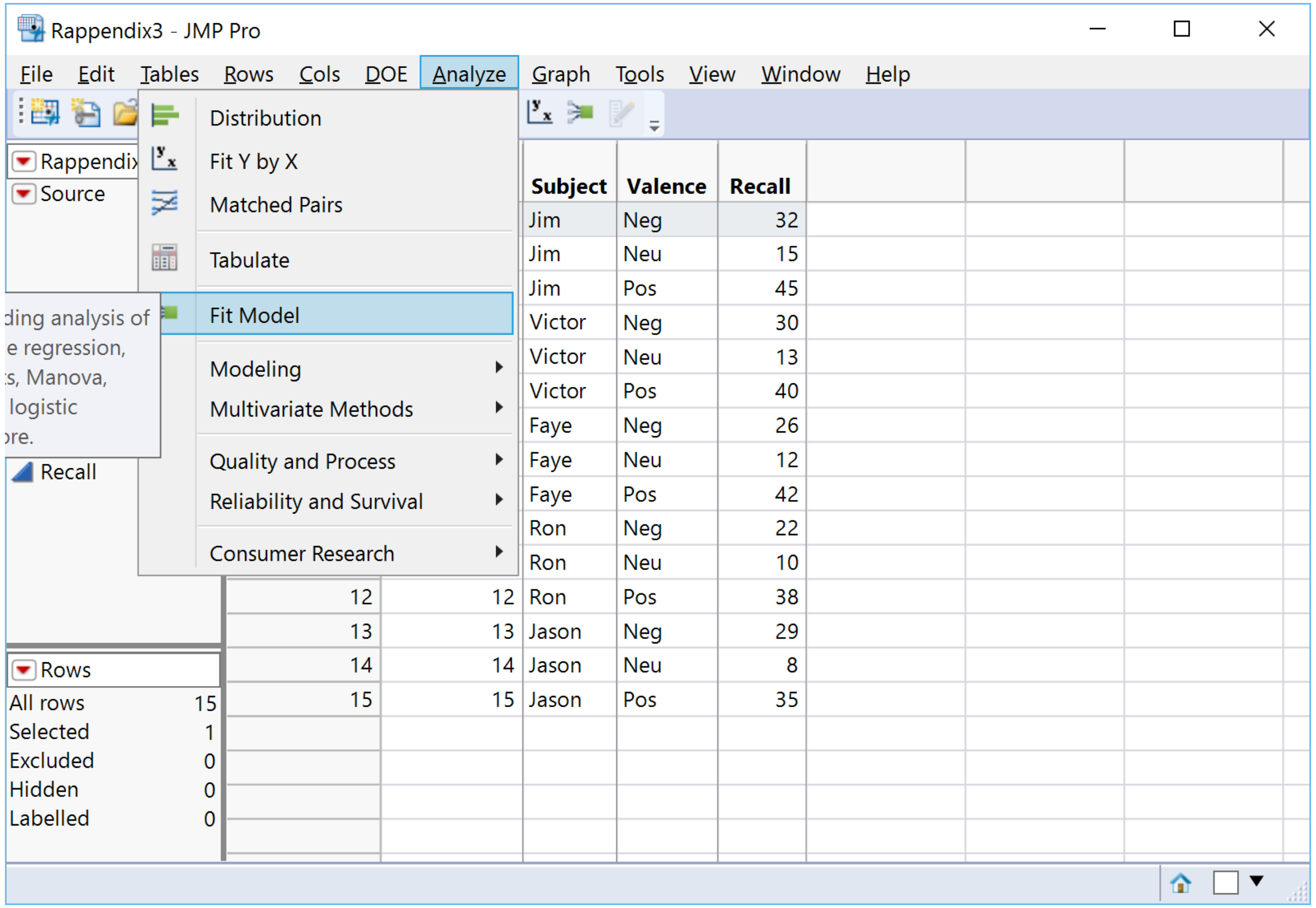
```
htests.model <- glht(linearM, linfct=mcp(W="Tukey"))
```

Within-Participants One-Way ANOVA in JMP

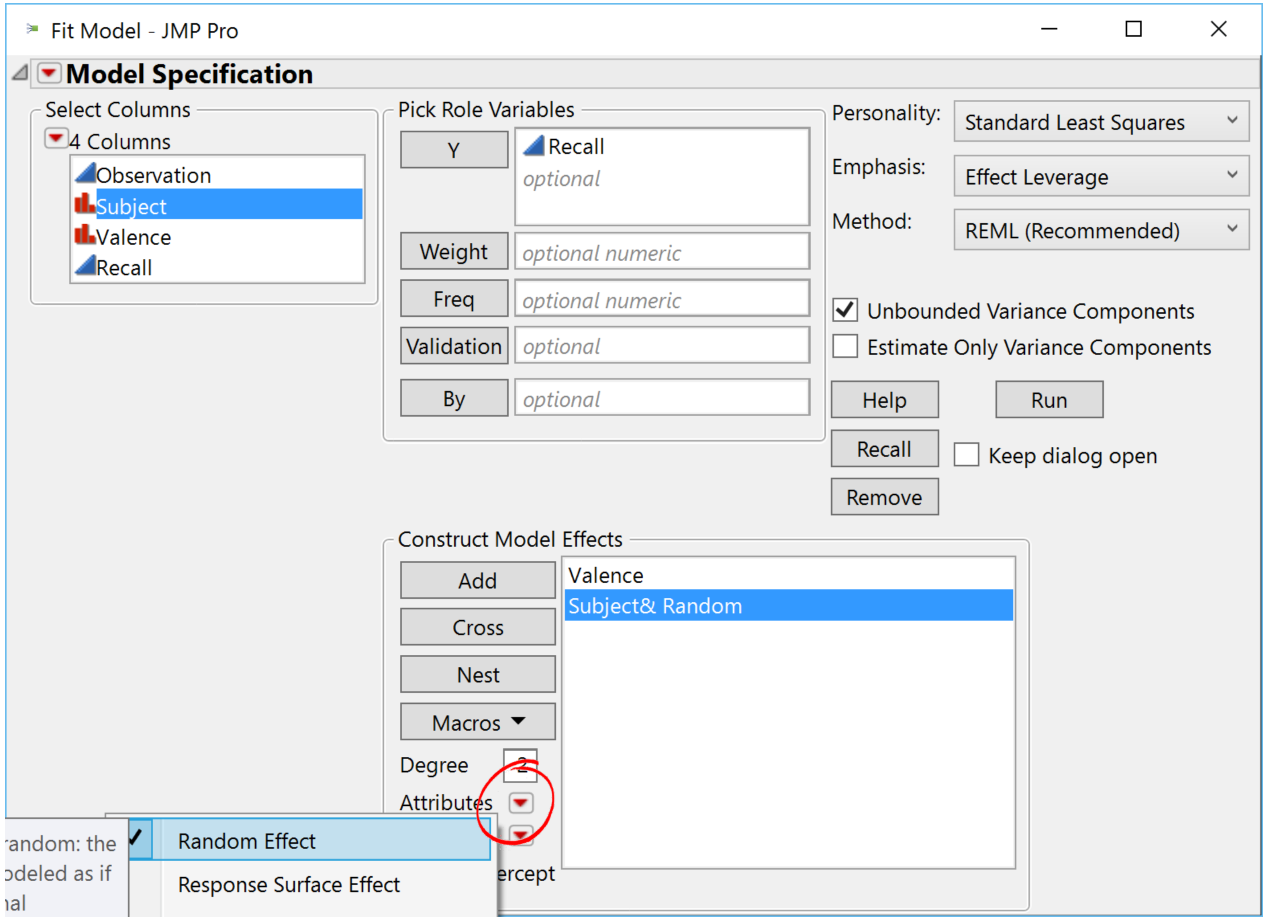


	Observation	Subject	Valence	Recall
1	1	Jim	Neg	32
2	2	Jim	Neu	15
3	3	Jim	Pos	45
4	4	Victor	Neg	30
5	5	Victor	Neu	13
6	6	Victor	Pos	40
7	7	Faye	Neg	26
8	8	Faye	Neu	12
9	9	Faye	Pos	42
10	10	Ron	Neg	22
11	11	Ron	Neu	10
12	12	Ron	Pos	38
13	13	Jason	Neg	29
14	14	Jason	Neu	8
15	15	Jason	Pos	35

Within-Participants One-Way ANOVA in JMP



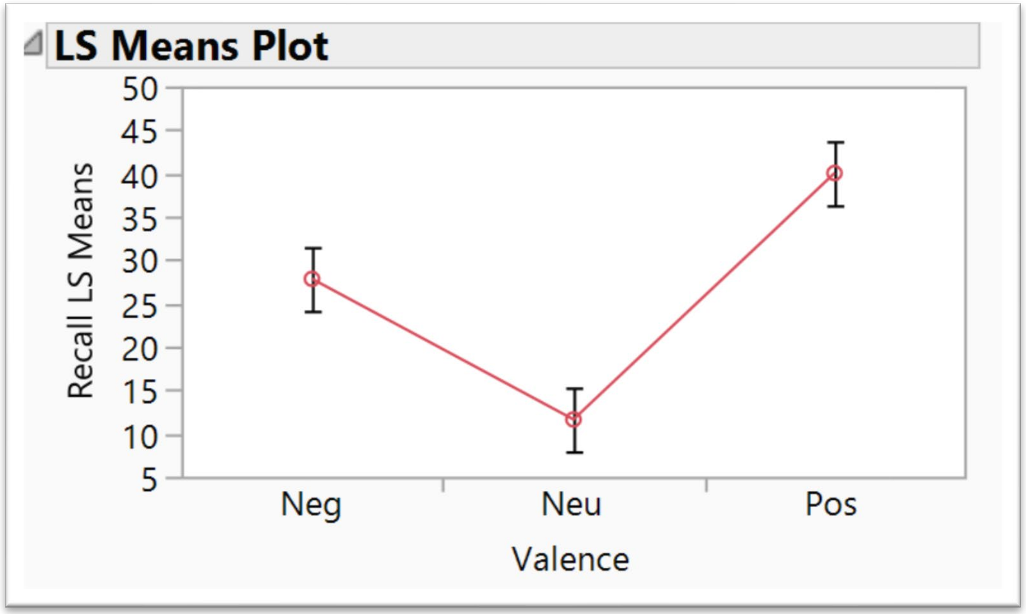
Within-Participants One-Way ANOVA in JMP

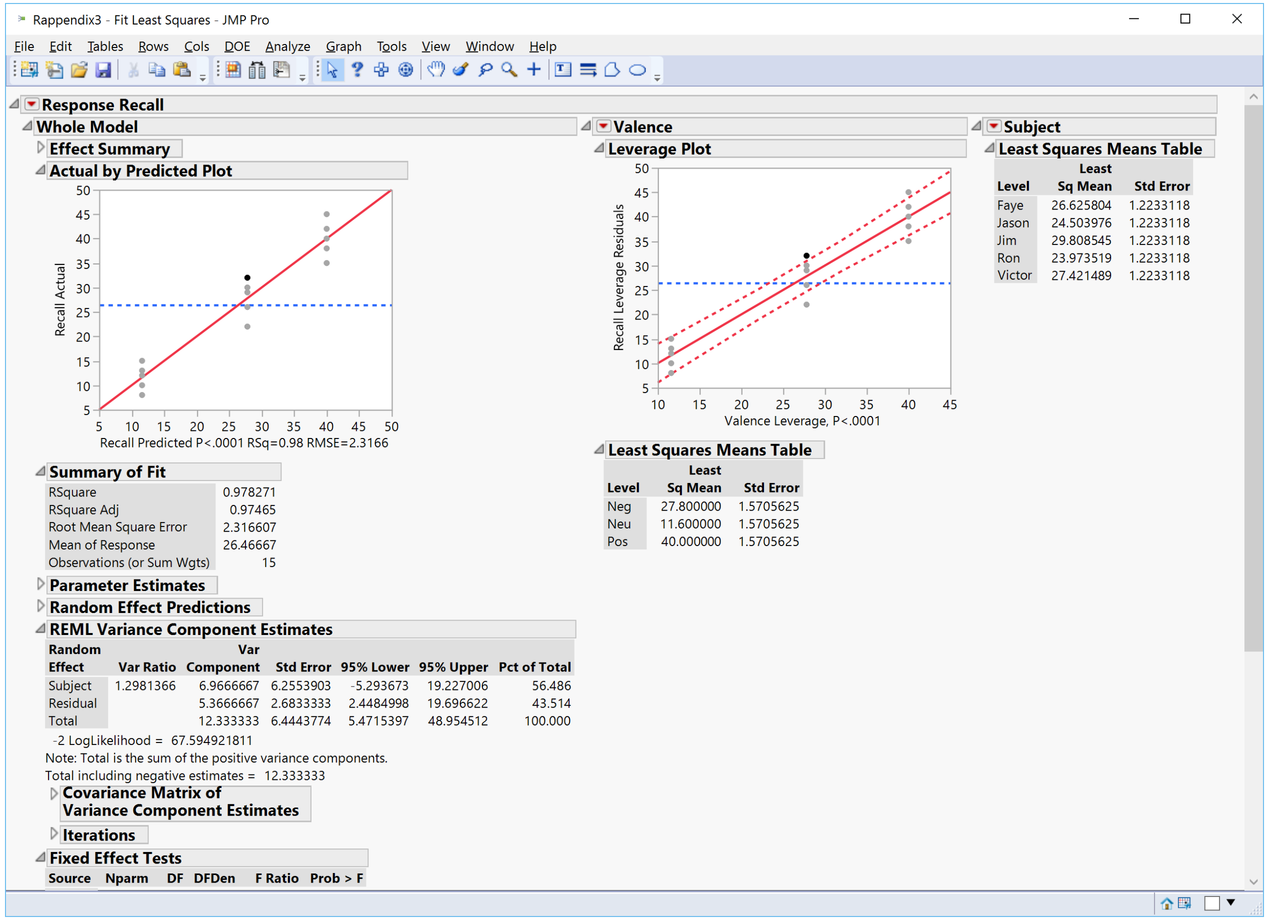


Within-Participants One-Way ANOVA in JMP

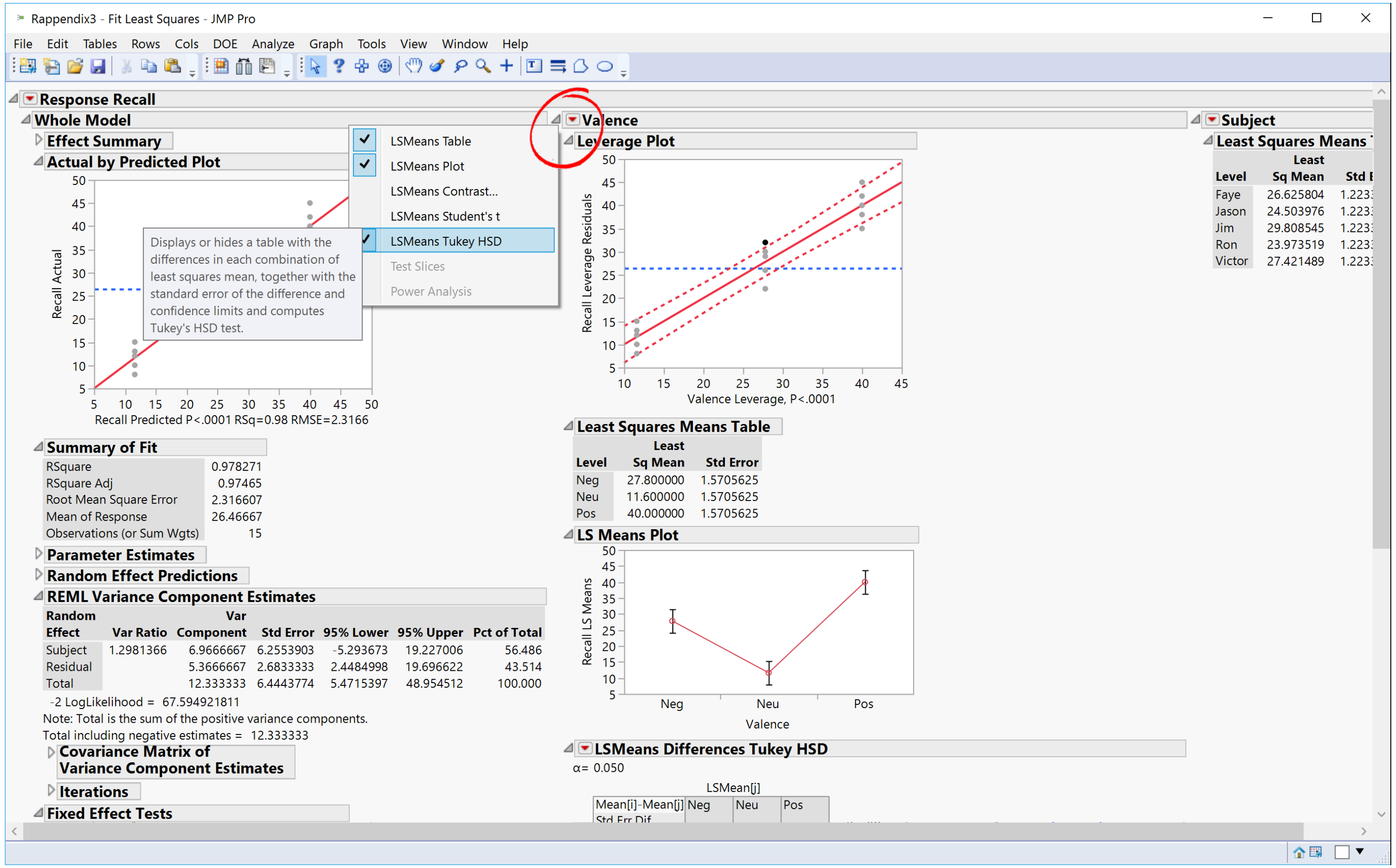
Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Valence	2	2	8	189.1056	<.0001*

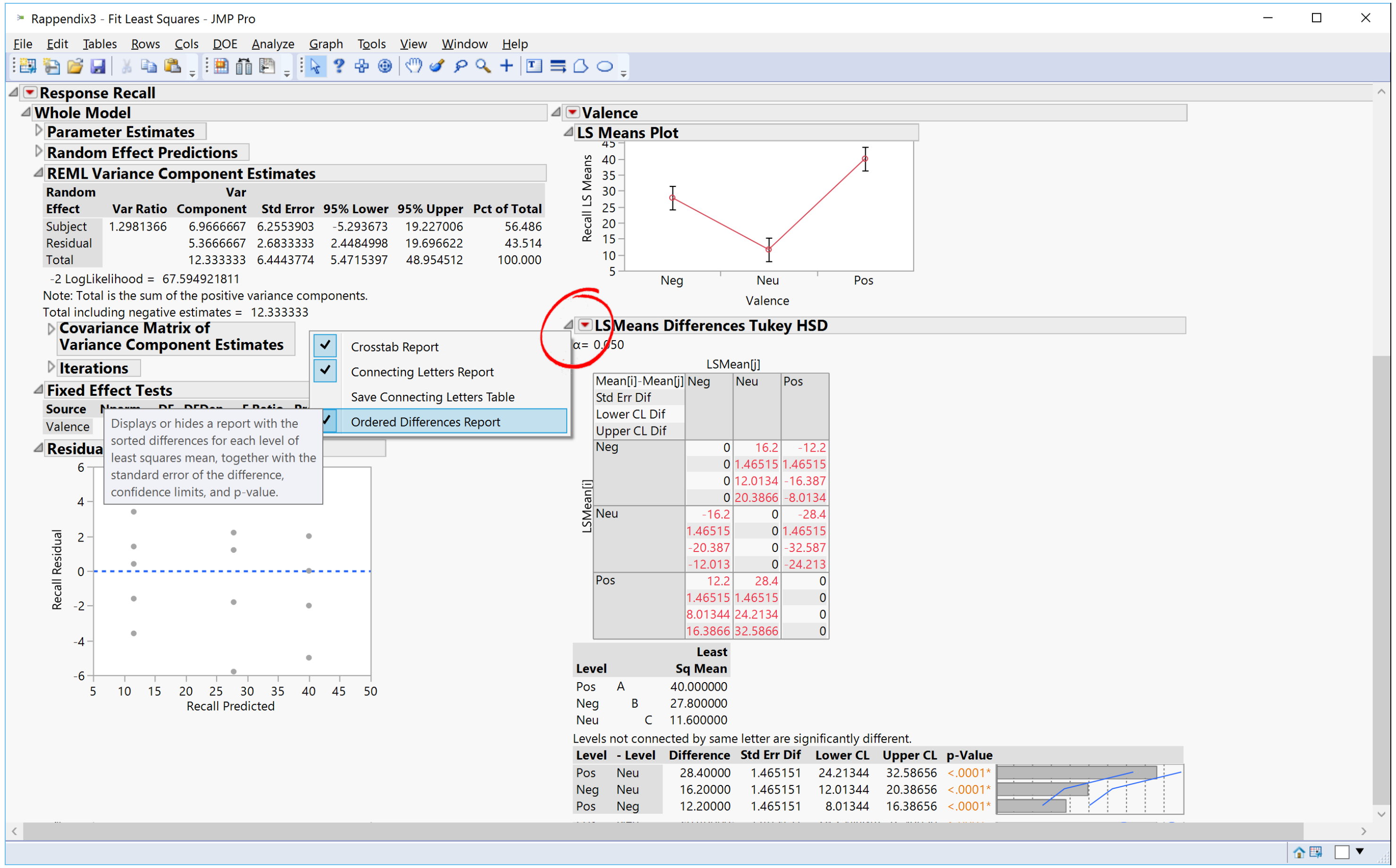




Post-Hoc Comparisons

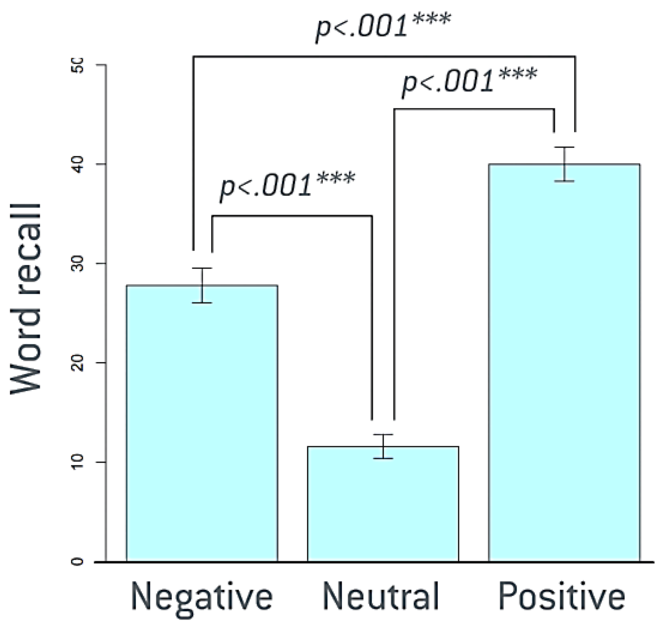


Post-Hoc Comparisons



Reporting Results

"We conducted a one-way analysis of variance (ANOVA) to test whether the valance of the words affected participants' recall of the words and found a significant effect of valance on word recall, $F(2,8) = 189.11$, $p < 0.001$. Post-hoc comparisons using Tukey's HSD test revealed differences between all pairwise tests ($p < .001$)."



Questions?

Types of ANOVA

- Different procedures
  - One-way ANOVA – single factor, two or more levels
    - Single factor, two level ANOVA is equivalent to a t-test
  - **Two-way ANOVA – two factors, two or more levels for each**
  - Multi-way ANOVA – multiple factors
- Different ways of modeling
  - Fixed effects, random effects, mixed effect

Two-way ANOVA

Two-Way Independent Measures ANOVA

For between-participants design

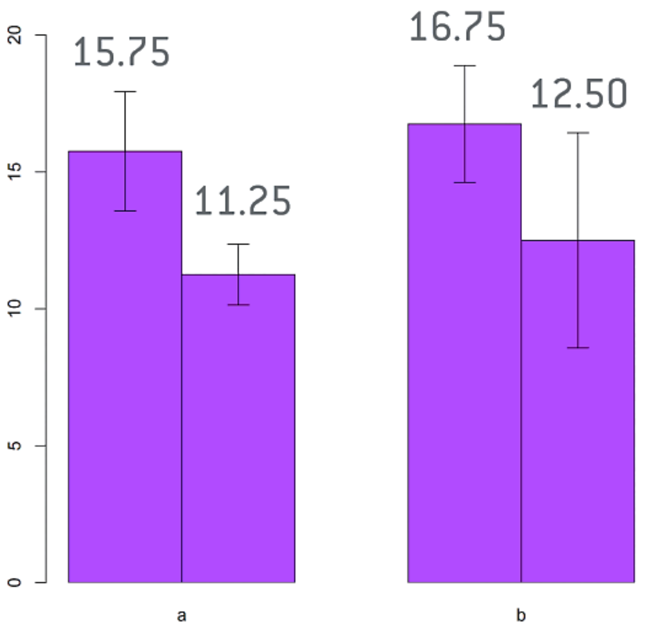
Example: Effect of dosage and gender on alertness

```
> data.tw.b=read.table("http://personality-project.org/r/datasets/R.appendix2.data",header=TRUE)

> fit.tw.b = aov(Alertness~Gender*Dosage,data=data.tw.b)

> summary(fit.tw.b)

               Df Sum Sq Mean Sq F value Pr(>F)

Gender          1  76.56   76.56   2.952  0.111

Dosage          1   5.06    5.06   0.195  0.666

Gender:Dosage   1   0.06    0.06   0.002  0.962

Residuals      12 311.25   25.94
```

fit <- aov(y~(B1*B2), data)

Two-Way Independent Measures ANOVA in R

```
> library(sciplot)

> bargraph.CI(data.tw.b$Dosage,
data.tw.b$Alertness, ylim=c(0,20),col="Purple",
group=data.tw.b$Gender)
```

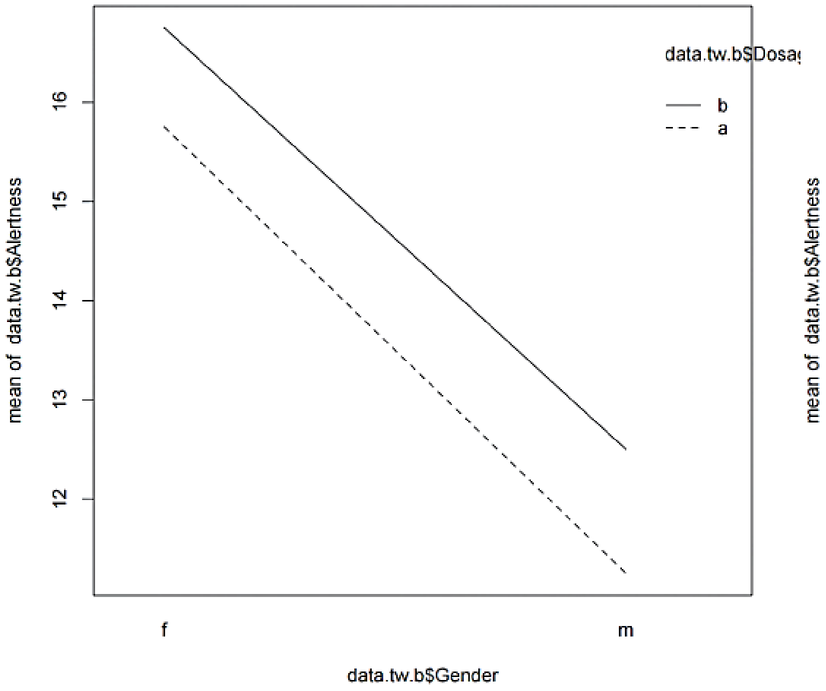


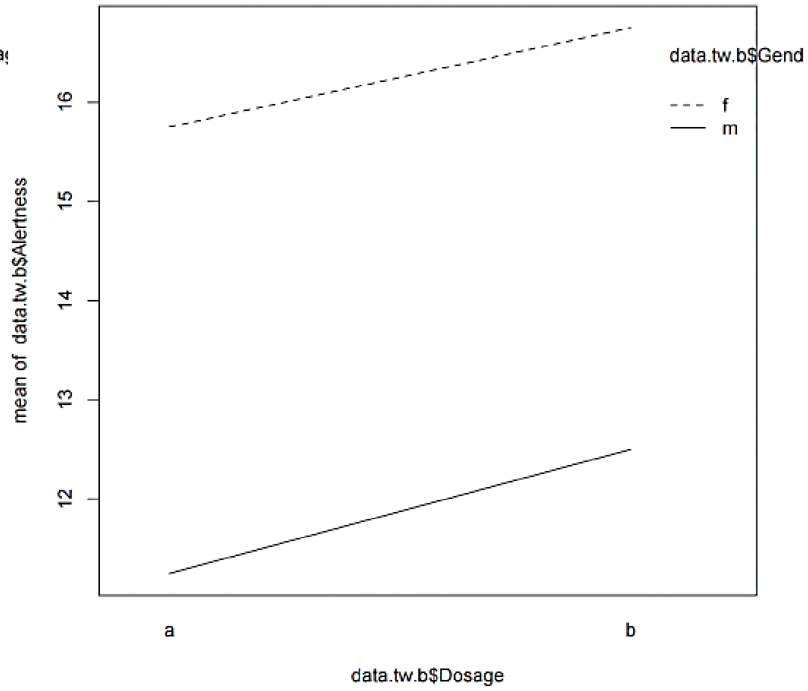
Test for interactions using linear model

```
> summary.lm(fit.tw.b)

Call:

aov(formula = Alertness ~ Gender * Dosage, data = data.tw.b)

Residuals:

   Min     1Q Median     3Q    Max

-6.500 -3.375  0.000  1.562 10.500

Coefficients:

                Estimate Std. Error t value Pr(>|t|)

(Intercept)       15.750      2.546   6.185 4.69e-05 ***

Genderm           -4.500      3.601  -1.250    0.235

Dosageb            1.000      3.601   0.278    0.786

Genderm:Dosageb    0.250      5.093   0.049    0.962

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.093 on 12 degrees of freedom

Multiple R-squared:  0.2079,      Adjusted R-squared:
0.009862

F-statistic:  1.05 on 3 and 12 DF,  p-value: 0.4062
```

Visualizing Interactions

```
> interaction.plot(data.tw.b$Gender,data.tw.b$Dosage,data.tw.b$Alertness)
> interaction.plot(data.tw.b$Dosage,data.tw.b$Gender,data.tw.b$Alertness)
```





Model Simplification

Simplifying your model by removing factors that don’t contribute to the model

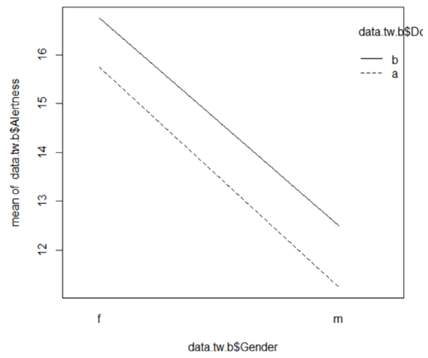
- Frees *df* providing power in your analysis
- Rule of thumb (Myers & Well, 1995)
  - Include all potentially relevant predictors initially and **consider** removing a term if:
    - You have *a priori* reason to think that it is not contributing **and**
    - The term is not significant at the .25 level

Simplifying the Model

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        15.750      2.546   6.185 4.69e-05 ***
Genderm            -4.500      3.601  -1.250    0.235
Dosageb             1.000      3.601   0.278    0.786
Genderm:Dosageb     0.250      5.093   0.049    0.962
```

Remove Gender x Dosage

```
> fit.tw.b =
aov(Alertness~Gender+Dosage,data=data.tw.b)
> summary(fit.tw.b)
```



Simplifying the Model

```
> summary(fit.tw.b)
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
Gender       1   76.56    76.56    3.197  0.0971 .
Dosage       1    5.06     5.06    0.211  0.6533
Residuals   13  311.31    23.95
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> fit.tw.b = aov(Alertness~Gender,data=data.tw.b)
> summary(fit.tw.b)
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
Gender       1    76.6    76.56    3.388   0.087 .
Residuals   14   316.4    22.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remove Dosage

What did we do wrong?

Simplifying the Model

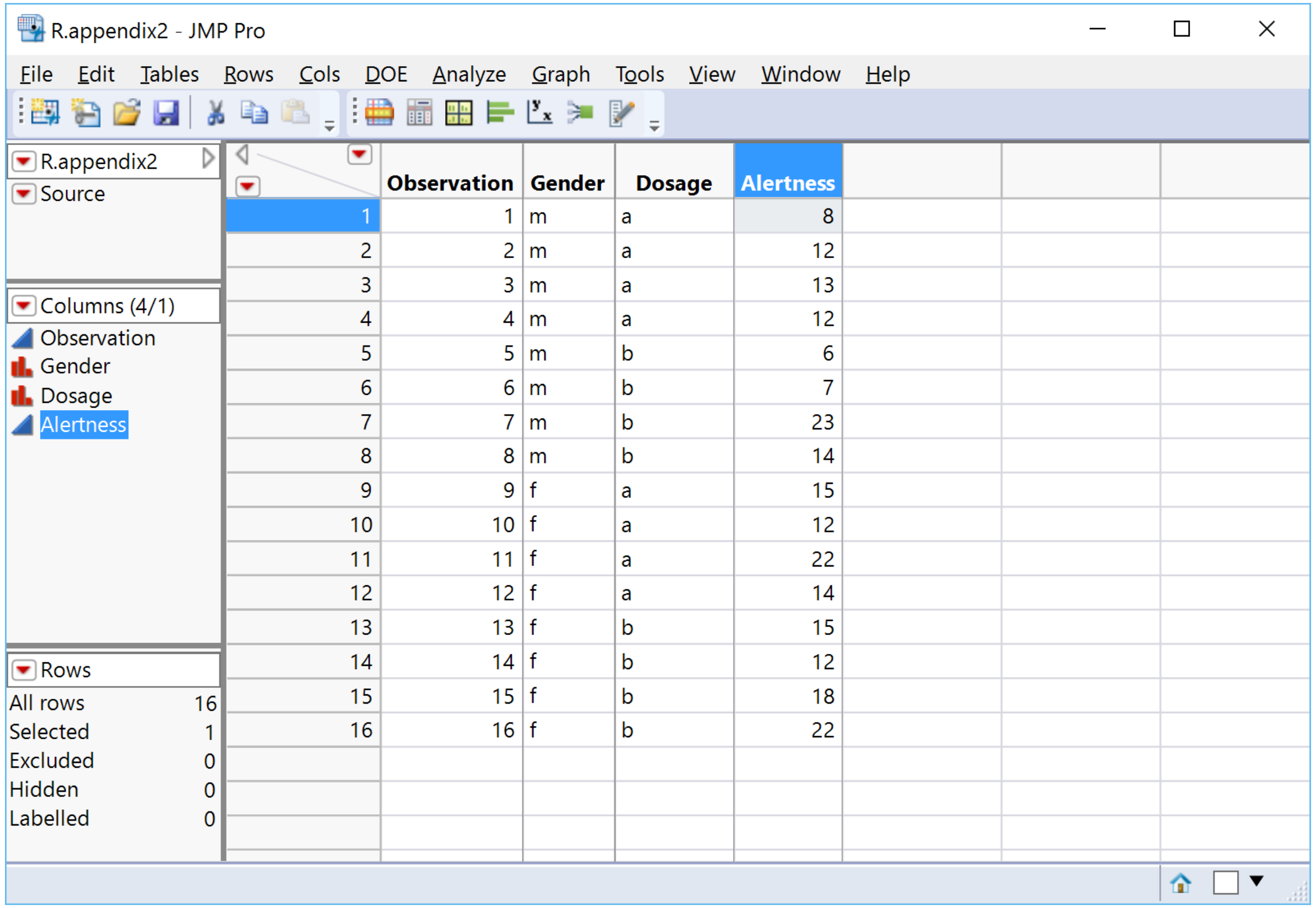
```
> summary(fit.tw.b)
            Df Sum Sq Mean Sq F value Pr(>F)
Gender       1  76.56   76.56   3.197 0.0971 .
Dosage       1   5.06    5.06   0.211 0.6533
Residuals   13 311.31   23.95
---
```

~~Remove Dosage~~

Only remove if >.25 AND *a priori* reason

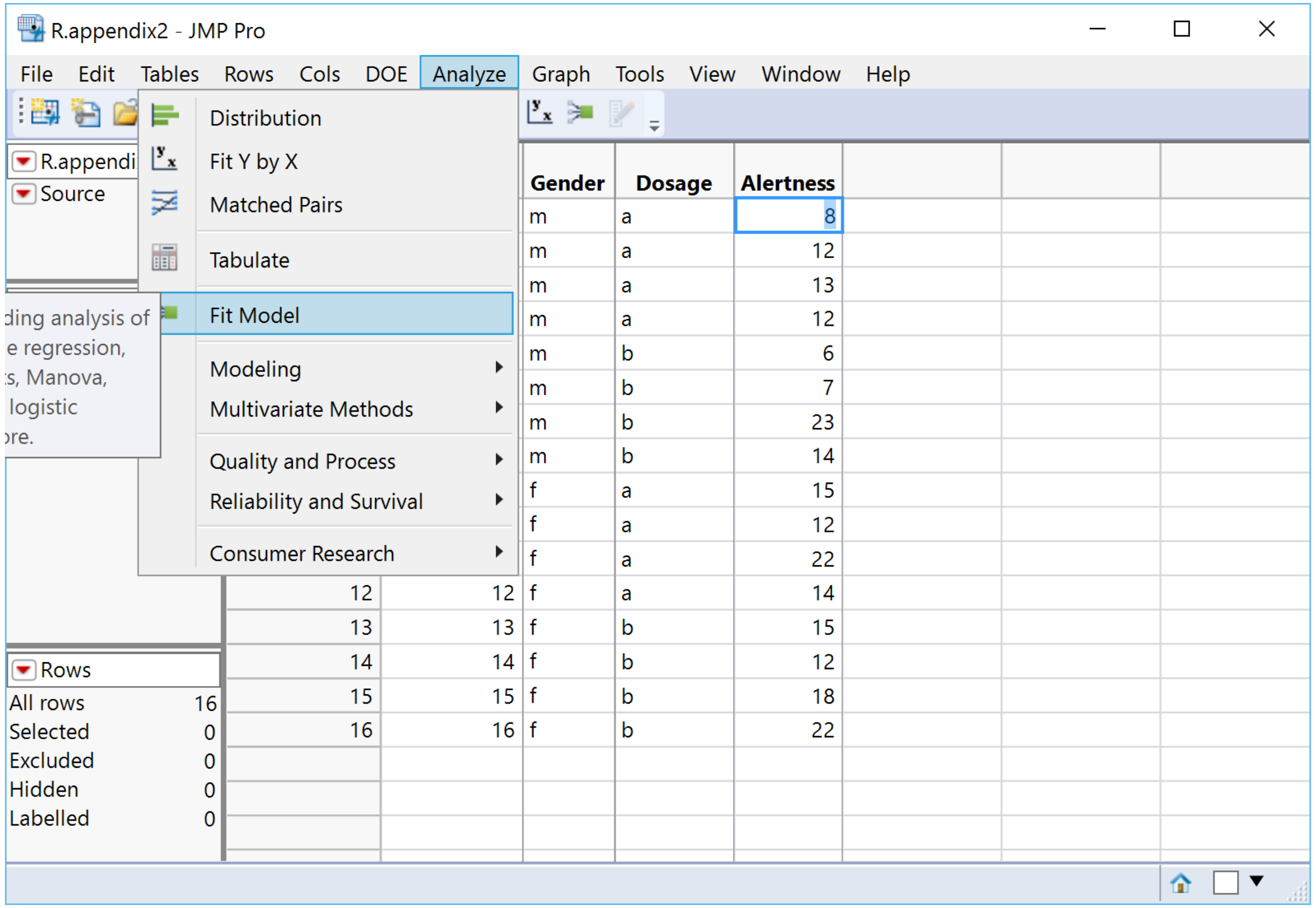
We hypothesized that dosage would affect alertness!

Two-Way Independent Measures ANOVA in JMP

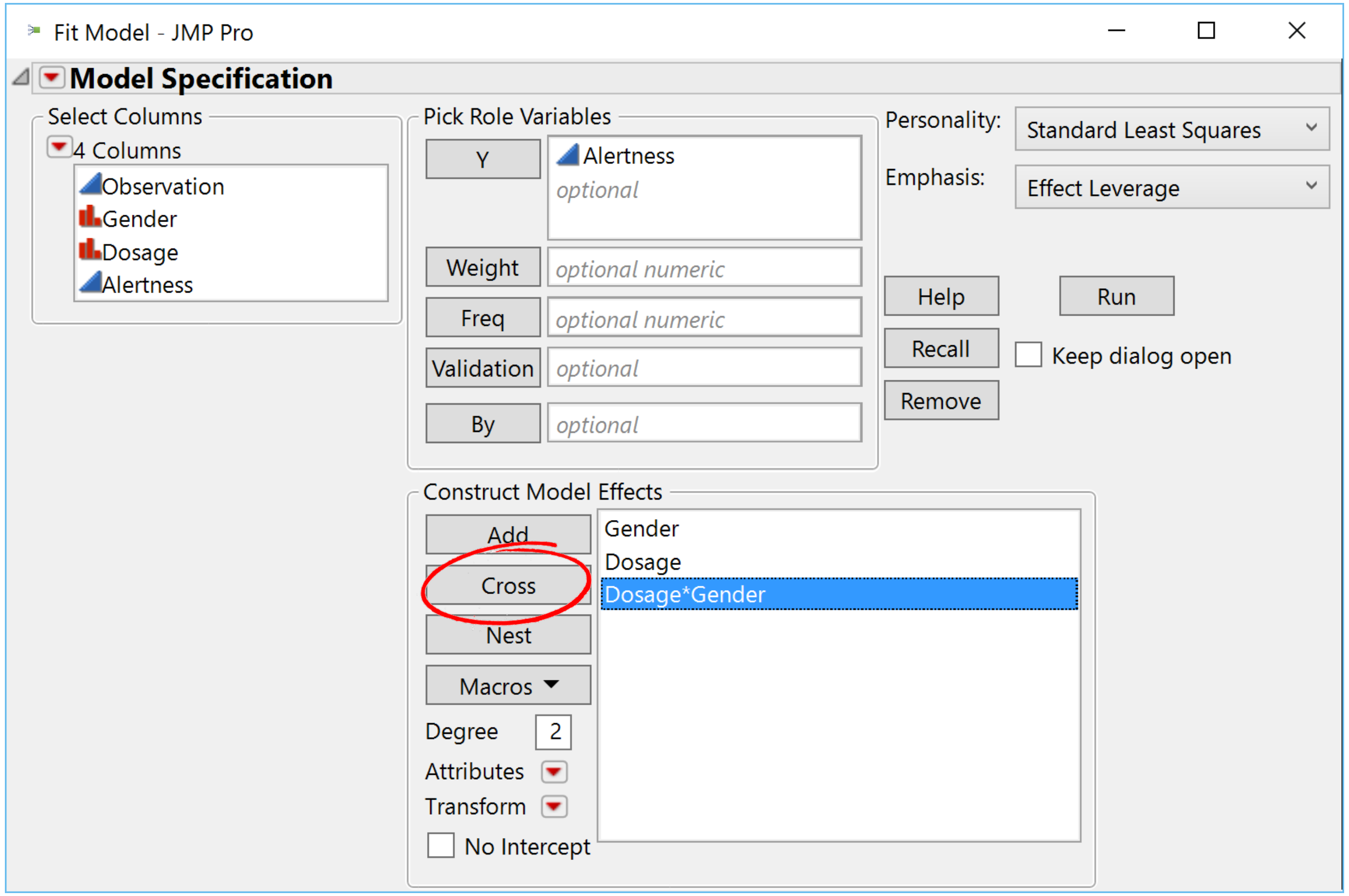


	Observation	Gender	Dosage	Alertness
1	1	m	a	8
2	2	m	a	12
3	3	m	a	13
4	4	m	a	12
5	5	m	b	6
6	6	m	b	7
7	7	m	b	23
8	8	m	b	14
9	9	f	a	15
10	10	f	a	12
11	11	f	a	22
12	12	f	a	14
13	13	f	b	15
14	14	f	b	12
15	15	f	b	18
16	16	f	b	22

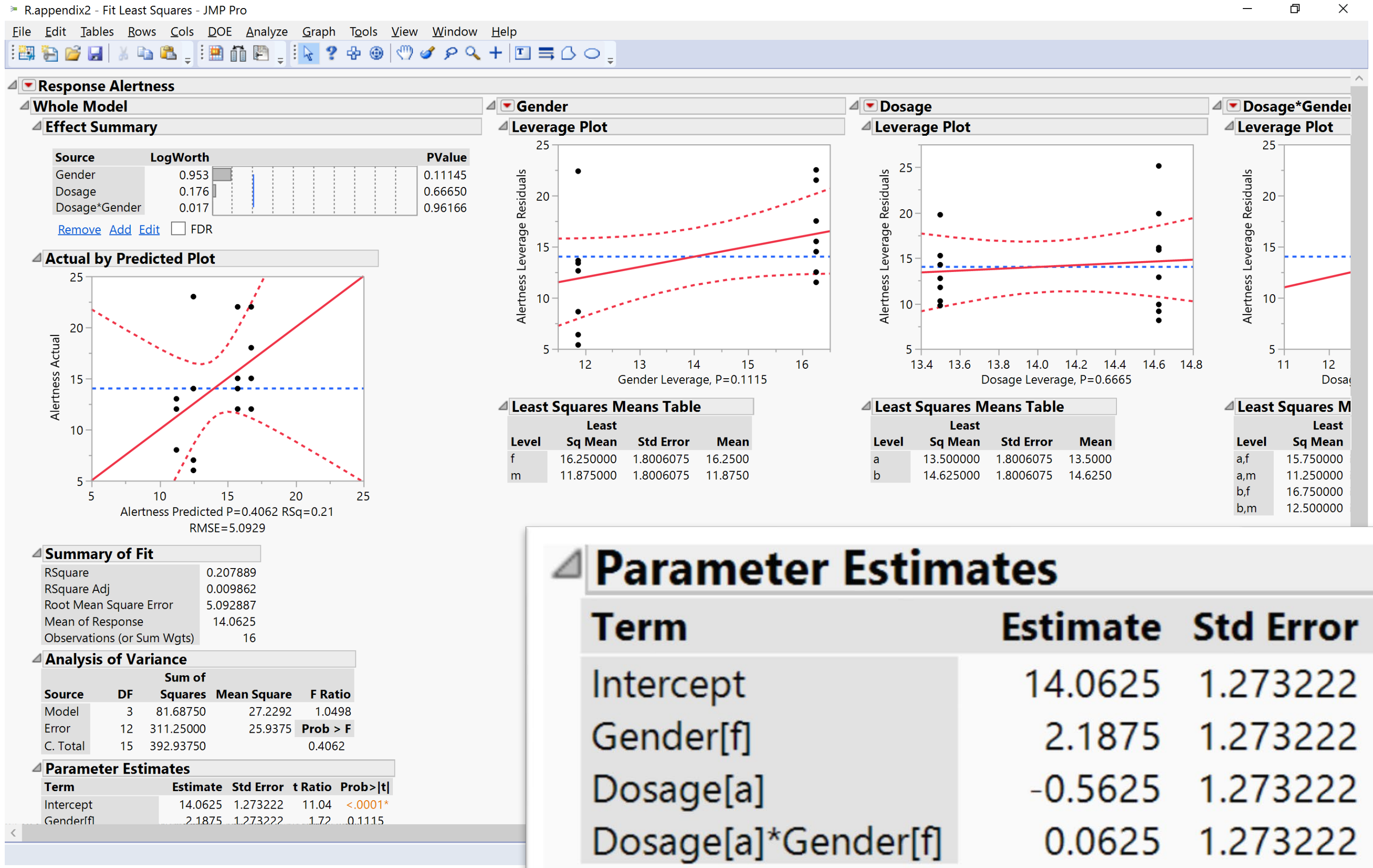
Two-Way Independent Measures ANOVA in JMP



Two-Way Independent Measures ANOVA in JMP



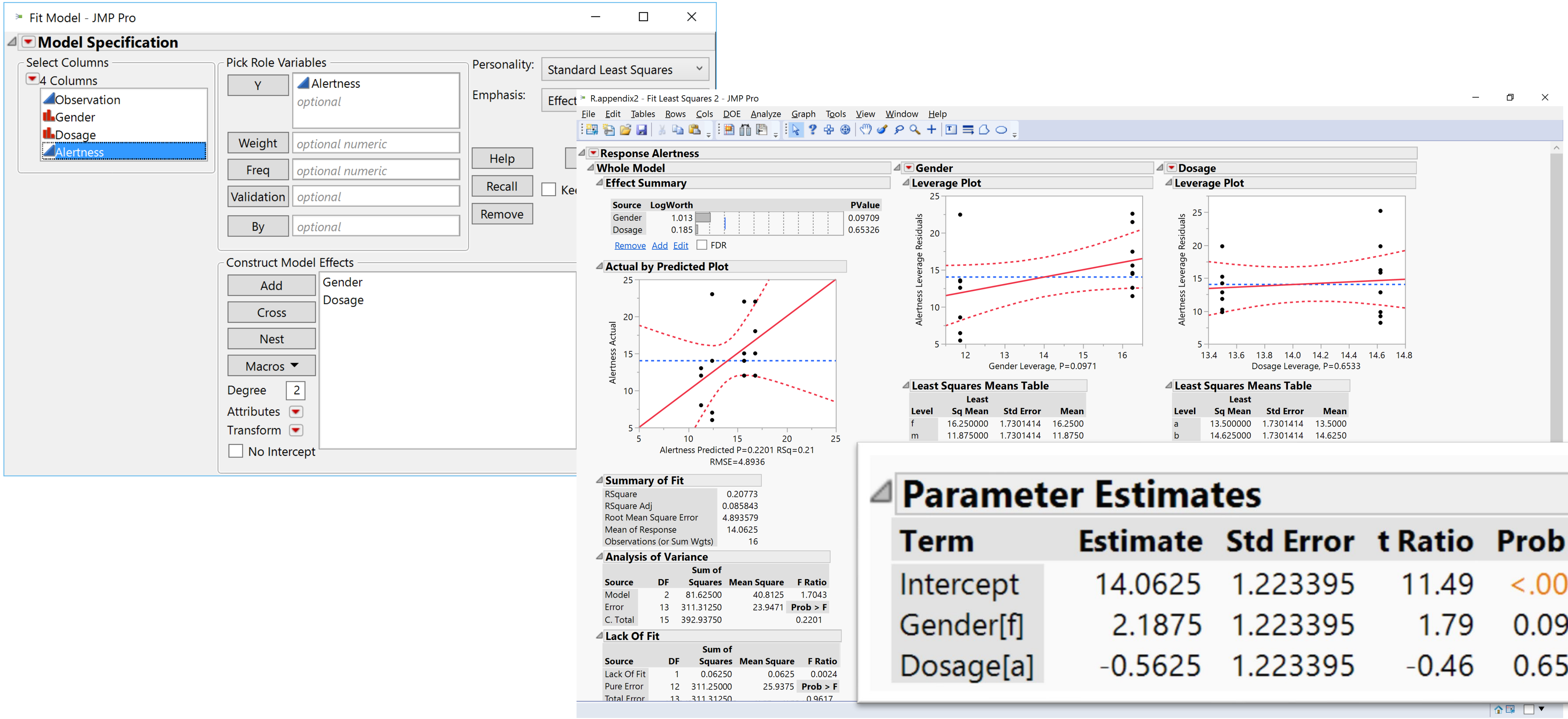
Two-Way Independent Measures ANOVA in JMP



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.0625	1.273222	11.04	<.0001*
Gender[f]	2.1875	1.273222	1.72	0.1115
Dosage[a]	-0.5625	1.273222	-0.44	0.6665
Dosage[a]*Gender[f]	0.0625	1.273222	0.05	0.9617

Simplified Model

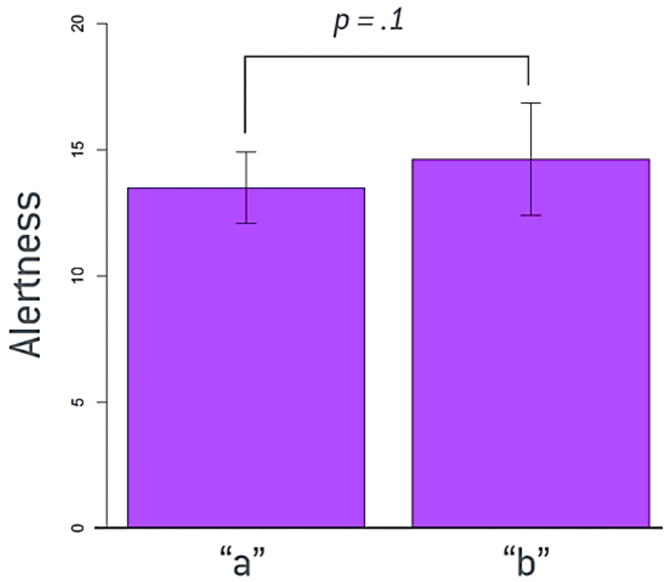


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.0625	1.223395	11.49	<.0001*
Gender[f]	2.1875	1.223395	1.79	0.0971
Dosage[a]	-0.5625	1.223395	-0.46	0.6533

Reporting Results

"We conducted a two-way analysis of variance (ANOVA) to test whether the dose of the drug and participant gender affected levels of alertness and found no significant main or interaction effects with the exception of a marginal main effect of dosage over alertness, $F(1,13) = 3.20$, $p = 0.1$."



Repeated Measures Two-Way ANVOA

Within-Participants Two-Way ANOVA in R

Example: Effect of valence and memory task on recall

```
> data.tw.w=read.table("http://personality-project.org/r/datasets/R.appendix4.data",header=TRUE)

> data.tw.w
```

	Observation	Subject	Task	Valence	Recall
1	1	Jim	Free	Neg	8
2	2	Jim	Free	Neu	9
3	3	Jim	Free	Pos	5
4	4	Jim	Cued	Neg	7
5	5	Jim	Cued	Neu	9
6	6	Jim	Cued	Pos	10
7	7	Victor	Free	Neg	12
8	8	Victor	Free	Neu	13
9	9	Victor	Free	Pos	14
10	10	Victor	Cued	Neg	16
11	11	Victor	Cued	Neu	13
12	12	Victor	Cued	Pos	14
13	13	Faye	Free	Neg	13
14	14	Faye	Free	Neu	13
15	15	Faye	Free	Pos	12
16	16	Faye	Cued	Neg	15
17	17	Faye	Cued	Neu	16
18	18	Faye	Cued	Pos	14
19	19	Ron	Free	Neg	12
20	20	Ron	Free	Neu	14
21	21	Ron	Free	Pos	15
22	22	Ron	Cued	Neg	17
23	23	Ron	Cued	Neu	18
24	24	Ron	Cued	Pos	20
25	25	Jason	Free	Neg	6
26	26	Jason	Free	Neu	7
27	27	Jason	Free	Pos	9
28	28	Jason	Cued	Neg	4
29	29	Jason	Cued	Neu	9
30	30	Jason	Cued	Pos	10

Within-Participants Two-Way ANOVA in R

```
> fit.tw.w =
aov(Recall~(Task*Valence)+Error(Subject/(Task*Valence)),data=data.tw.w)

> summary(fit.tw.w)
```

fit <- aov(y~(W1*W2)+Error(Subject/(W1*W2)), data)

```
Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4  349.1   87.28
Error: Subject:Task
          Df Sum Sq Mean Sq F value Pr(>F)
Task       1  30.00  30.000   7.347 0.0535
.
Residuals  4  16.33   4.083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1
Error: Subject:Valence
          Df Sum Sq Mean Sq F value Pr(>F)
Valence    2   9.80   4.900   1.459  0.288
Residuals  8  26.87   3.358
Error: Subject:Task:Valence
             Df Sum Sq Mean Sq F value
Pr(>F)
Task:Valence  2   1.40   0.700   0.291
0.755
Residuals     8  19.27   2.408
```

Simplify the Model

```
Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4  349.1   87.28
Error: Subject:Task
          Df Sum Sq Mean Sq F value Pr(>F)
Task       1  30.00  30.000   7.347 0.0535 .
Residuals  4  16.33   4.083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Error: Subject:Valence
          Df Sum Sq Mean Sq F value Pr(>F)
Valence    2   9.80   4.900   1.459  0.288
Residuals  8  26.87   3.358
Error: Subject:Task:Valence
             Df Sum Sq Mean Sq F value Pr(>F)
Task:Valence  2   1.40   0.700   0.291  0.755
Residuals     8  19.27   2.408
```

Remove Task x Valance

Simplify the Model

```
> fit.tw.w = aov(Recall~(Task+Valence)+Error(Subject/(Task+Valence)),data=data.tw.w)

> summary(fit.tw.w)

Error: Subject

          Df Sum Sq Mean Sq F value Pr(>F)

Residuals  4  349.1   87.28

Error: Subject:Task

          Df Sum Sq Mean Sq F value Pr(>F)

Task       1  30.00  30.000   7.347 0.0535 .

Residuals  4  16.33   4.083

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Subject:Valence

          Df Sum Sq Mean Sq F value Pr(>F)

Valence    2   9.80   4.900   1.459  0.288

Residuals  8  26.87   3.358

Error: Within

          Df Sum Sq Mean Sq F value Pr(>F)

Residuals 10  20.67   2.067
```

Remove Valance

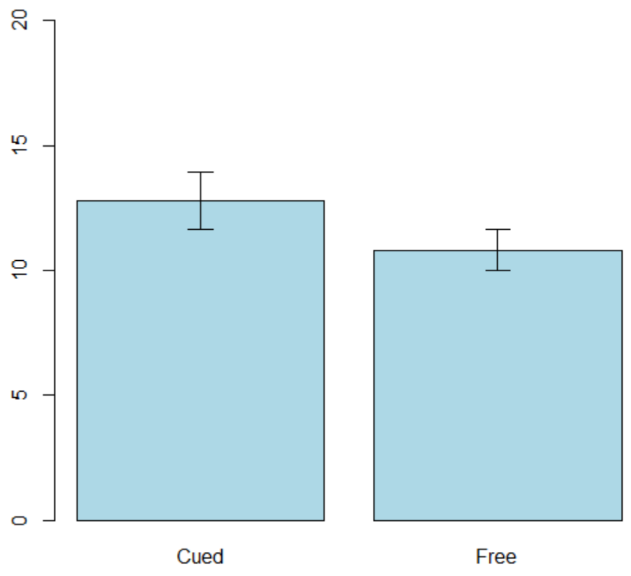
Simplify the Model

```
> fit.tw.w = aov(Recall~Task+Error(Subject/Task),data=data.tw.w)
> summary(fit.tw.w)
Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4  349.1   87.28
Error: Subject:Task
          Df Sum Sq Mean Sq F value Pr(>F)
Task       1  30.00  30.000   7.347 0.0535 .
Residuals  4  16.33   4.083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 20  57.33   2.867
```

Simplification is not necessary in a within-participants analysis!

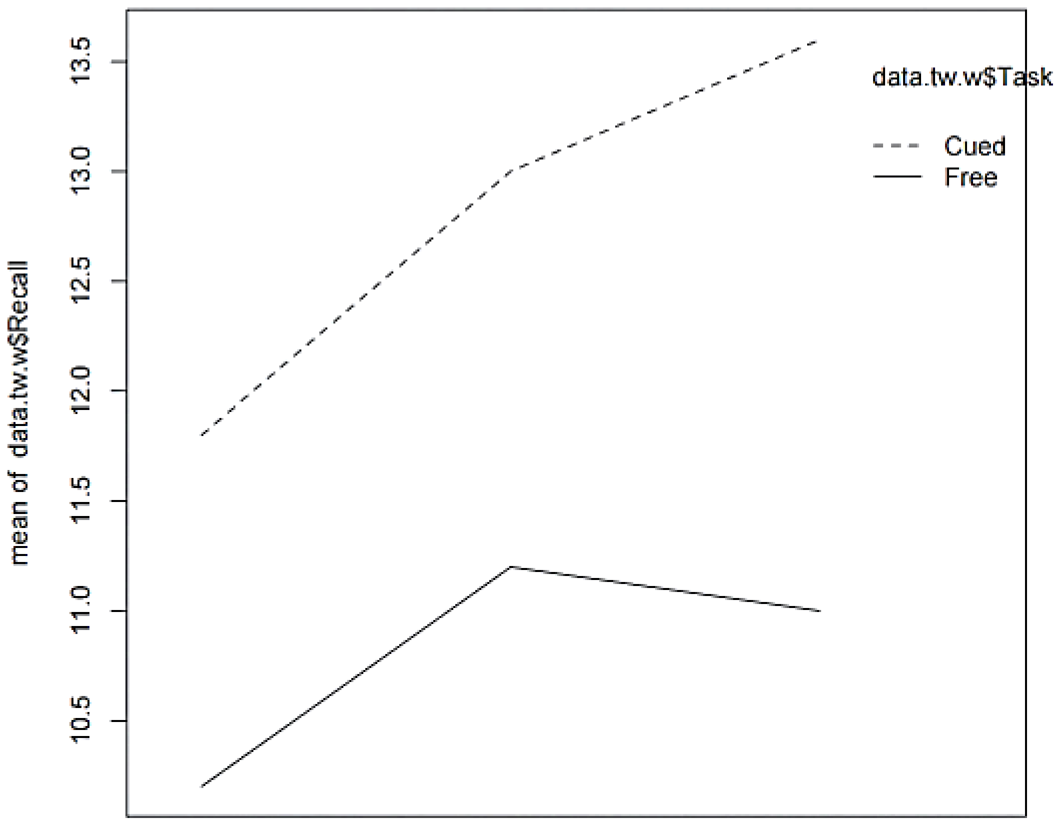
Visualizing Repeated Measures Two-Way ANOVA

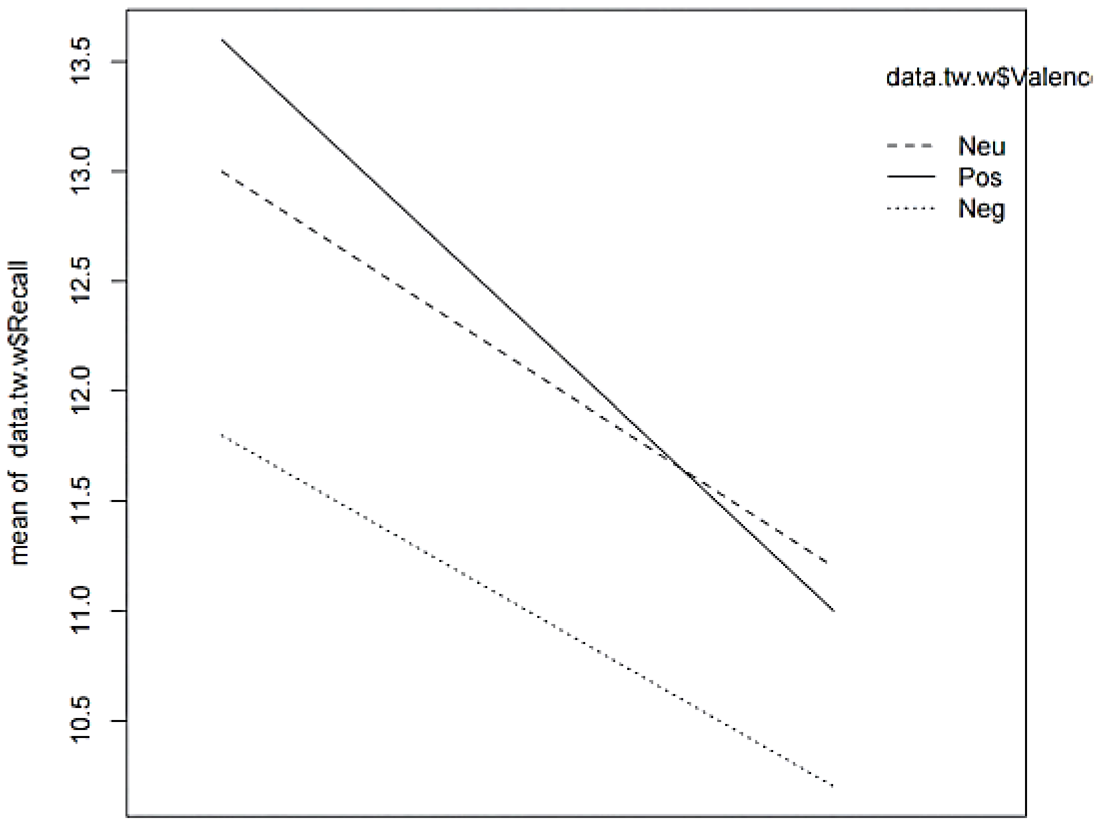
```
> library(sciplot)

> bargraph.CI(data.tw.w$Task, data.tw.w$Recall,
ylim=c(0,20), col="LightBlue")
```



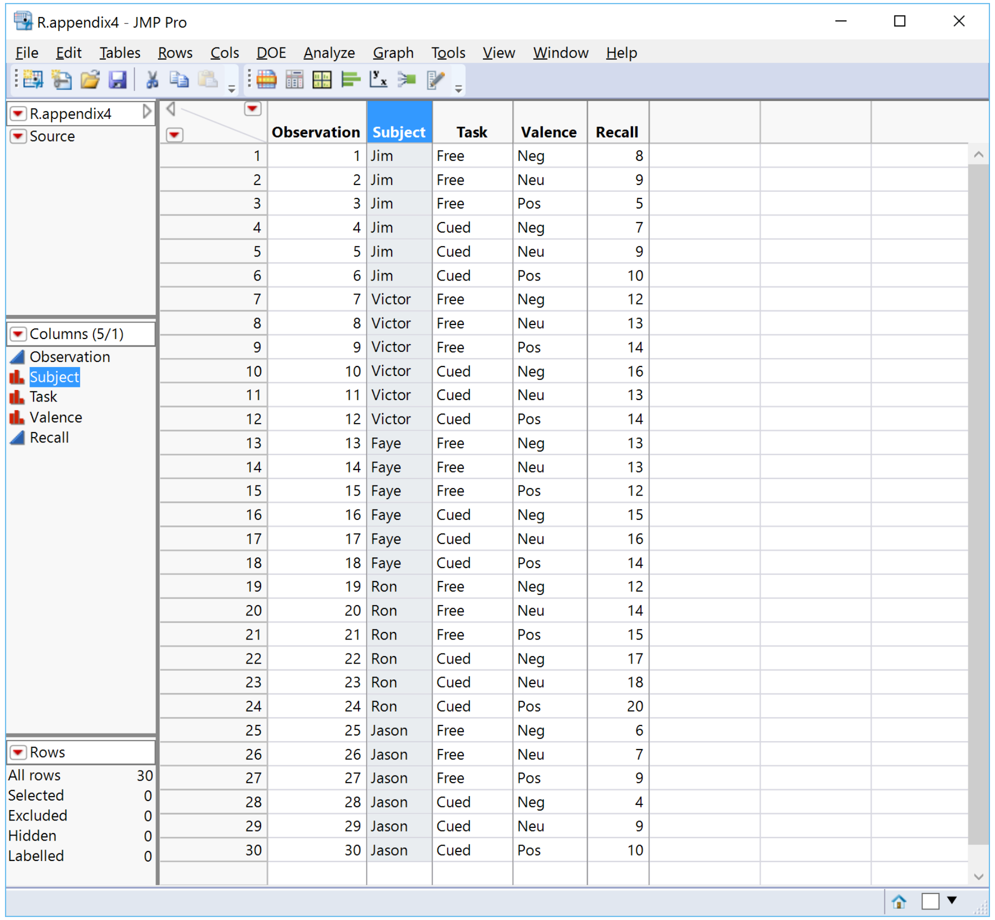
Visualizing Interactions

```
> interaction.plot(data.tw.w$Valence,data.tw.w$Task,data.tw.w$Recall)
> interaction.plot(data.tw.w$Task,data.tw.w$Valence,data.tw.w$Recall)
```



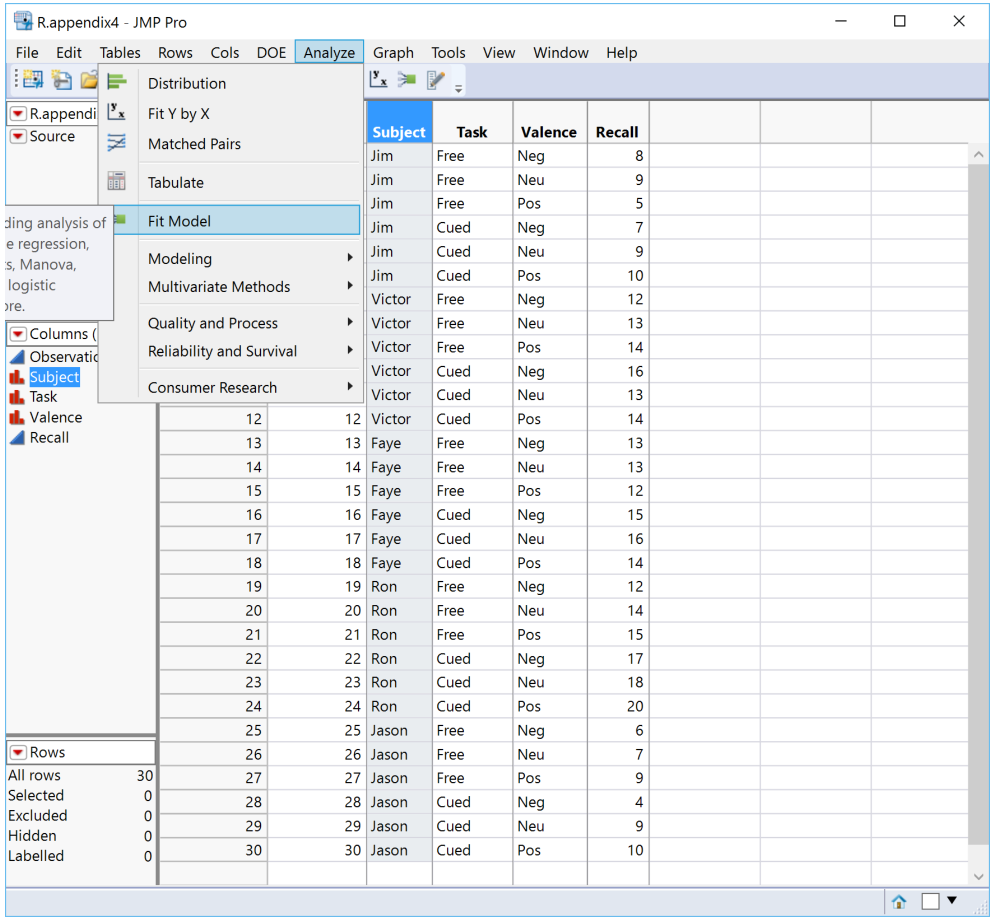


Repeated-Measures Two-Way ANOVA in JMP

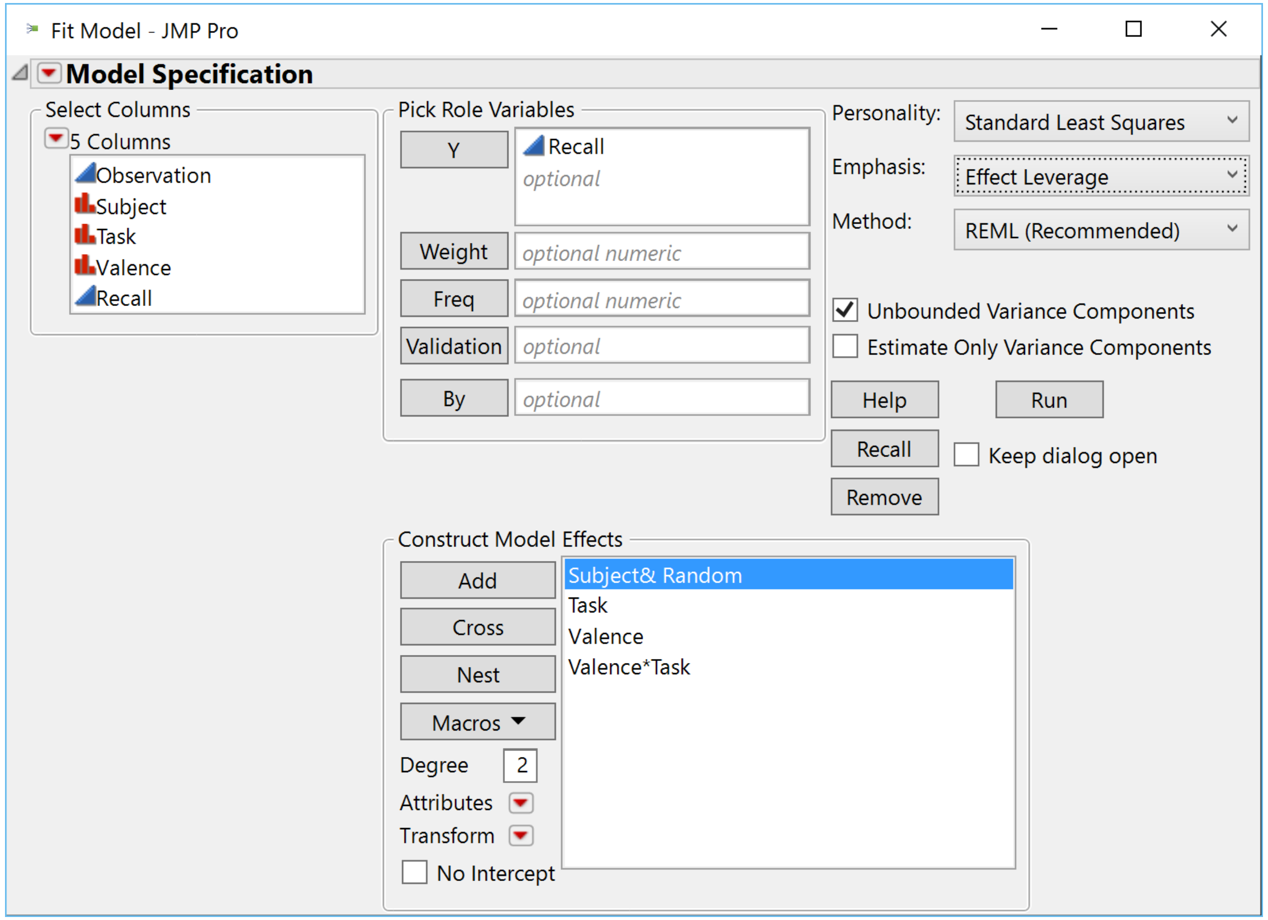


	Observation	Subject	Task	Valence	Recall
1	1	Jim	Free	Neg	8
2	2	Jim	Free	Neu	9
3	3	Jim	Free	Pos	5
4	4	Jim	Cued	Neg	7
5	5	Jim	Cued	Neu	9
6	6	Jim	Cued	Pos	10
7	7	Victor	Free	Neg	12
8	8	Victor	Free	Neu	13
9	9	Victor	Free	Pos	14
10	10	Victor	Cued	Neg	16
11	11	Victor	Cued	Neu	13
12	12	Victor	Cued	Pos	14
13	13	Faye	Free	Neg	13
14	14	Faye	Free	Neu	13
15	15	Faye	Free	Pos	12
16	16	Faye	Cued	Neg	15
17	17	Faye	Cued	Neu	16
18	18	Faye	Cued	Pos	14
19	19	Ron	Free	Neg	12
20	20	Ron	Free	Neu	14
21	21	Ron	Free	Pos	15
22	22	Ron	Cued	Neg	17
23	23	Ron	Cued	Neu	18
24	24	Ron	Cued	Pos	20
25	25	Jason	Free	Neg	6
26	26	Jason	Free	Neu	7
27	27	Jason	Free	Pos	9
28	28	Jason	Cued	Neg	4
29	29	Jason	Cued	Neu	9
30	30	Jason	Cued	Pos	10

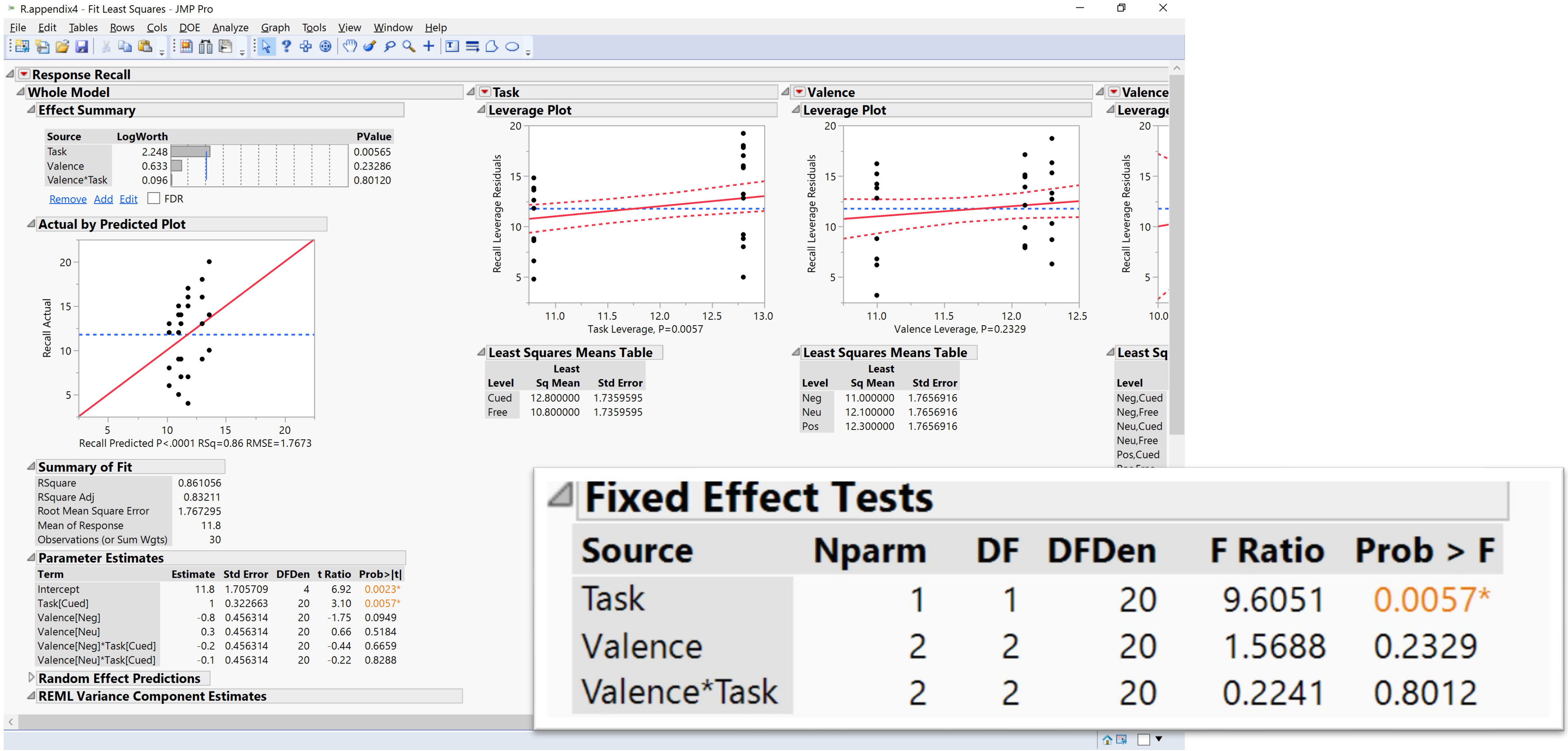
Repeated-Measures Two-Way ANOVA in JMP



Repeated-Measures Two-Way ANOVA in JMP



Repeated-Measures Two-Way ANOVA in JMP



Source	LogWorth	PValue
Task	2.248	0.00565
Valence	0.633	0.23286
Valence*Task	0.096	0.80120

Summary of Fit

RSquare	0.861056
RSquare Adj	0.83211
Root Mean Square Error	1.767295
Mean of Response	11.8
Observations (or Sum Wgts)	30

Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	11.8	1.705709	4	6.92	0.0023*
Task[Cued]	1	0.322663	20	3.10	0.0057*
Valence[Neg]	-0.8	0.456314	20	-1.75	0.0949
Valence[Neu]	0.3	0.456314	20	0.66	0.5184
Valence[Neg]*Task[Cued]	-0.2	0.456314	20	-0.44	0.6659
Valence[Neu]*Task[Cued]	-0.1	0.456314	20	-0.22	0.8288

Least Squares Means Table

Level	Least Sq Mean	Std Error
Cued	12.800000	1.7359595
Free	10.800000	1.7359595

Least Squares Means Table

Level	Least Sq Mean	Std Error
Neg	11.000000	1.7656916
Neu	12.100000	1.7656916
Pos	12.300000	1.7656916

Least Sq

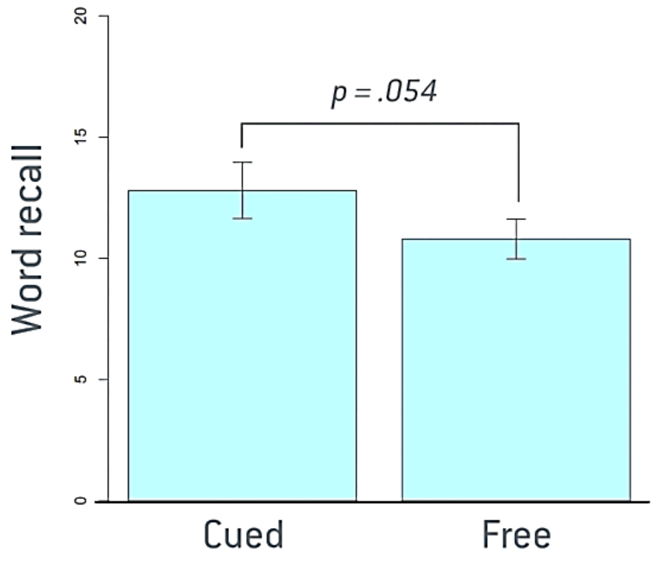
Level
Neg,Cued
Neg,Free
Neu,Cued
Neu,Free
Pos,Cued

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Task	1	1	20	9.6051	0.0057*
Valence	2	2	20	1.5688	0.2329
Valence*Task	2	2	20	0.2241	0.8012

Reporting Results

"We conducted a two-way analysis of variance (ANOVA) to test whether the memory task and the valence of the words affected participants' recall of the words and found no significant main or interaction effects, except a marginal effect of memory task on word recall F(1,22) = 7.35, p = 0.06."



Questions?

Types of ANOVA

- Different procedures
  - One-way ANOVA – single factor, two or more levels
    - Single factor, two level ANOVA is equivalent to a t-test
  - Two-way ANOVA – two factors, two or more levels for each
  - **Multi-way ANOVA – multiple factors**
- Different ways of modeling
  - Fixed effects, random effects, mixed effect

Mixed-Model ANOVA

Mixed-Model ANOVA in R

Within- and between-participants comparisons

Example: effect of drug dosage (between) and memory task (within) on recall

```
> data.mm=read.table("http://personality-project.org/r/datasets/R.appendix5.data",header=TRUE)

> fit.mm = aov(Recall~(Task*Dosage)+Error(Subject/Task)+Dosage,data=data.mm)
```

```
fit <- aov(y~(W*B)+Error(Subject/(W))+B,data)
```

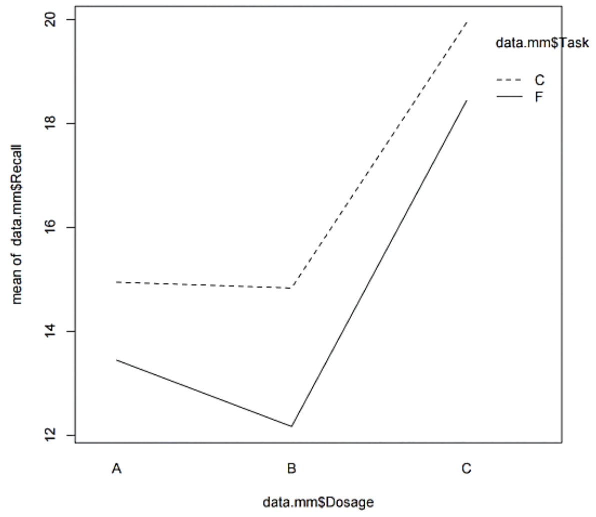
Mixed-Model ANOVA in R

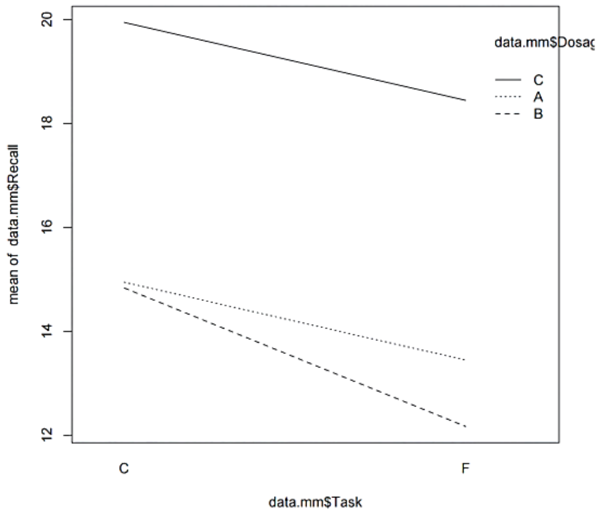
```
> summary(fit.mm)

Error: Subject

          Df Sum Sq Mean Sq F value Pr(>F)

Dosage     2  694.9   347.5   2.965 0.0822 .

Residuals 15 1757.6   117.2

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Error: Subject:Task

            Df  Sum Sq Mean Sq F value   Pr(>F)

Task         1  96.33   96.33  43.134 8.92e-06 ***

Task:Dosage  2   8.17    4.08   1.828    0.195

Residuals   15  33.50    2.23

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Error: Within

          Df Sum Sq Mean Sq F value Pr(>F)

Residuals 72  160.7   2.232
```

Interactions

Dosage and Gender

```
> interaction.plot(data.mm$Dosage,data.mm$Task,data.mm$Recall)
> interaction.plot(data.mm$Task,data.mm$Dosage,data.mm$Recall)
```





Post-Hoc Results

Tedious, unfortunately

- We need to compute the F value manually and correct for the **Type 1 error**
- To compare “A” and “B”:

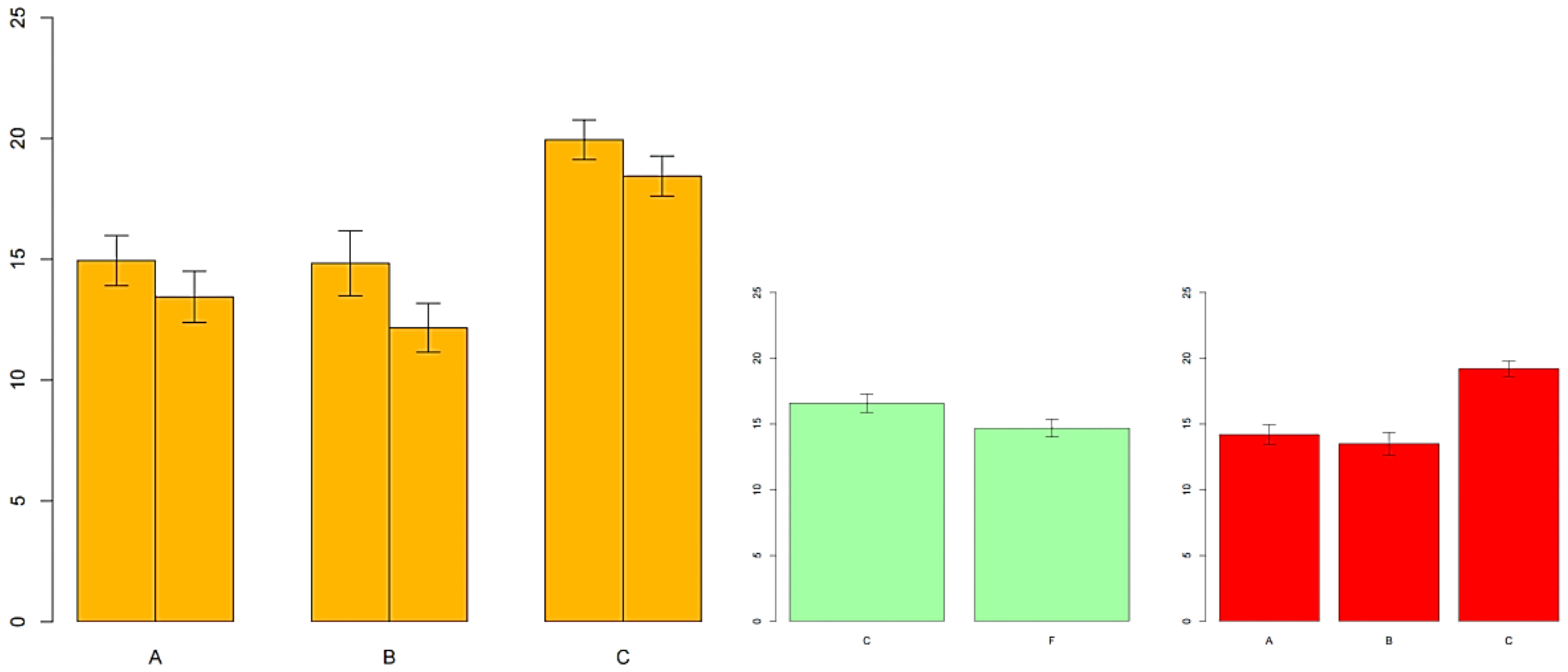
```
> SScomp <-
(mean(data.mm$Recall[data.mm$Dosage=="A"])-
mean(data.mm$Recall[data.mm$Dosage=="B"]))^2
> dfcomp <- 1
> n <- 36 # participants per group
> SSerr <- 1757.61 # residual from Dosage main
effect test
> dferr <- 15 # df from Dosage main effect test
> MSerr <- SSerr / dferr
> Fcomp <- (n * SScomp/2)/MSerr
> pcomp <- 1-pf(Fcomp, dfcomp, dferr)
> qobs <- sqrt(2*Fcomp)
> ngroups <- length(levels(data.mm$Dosage))
> ptk <- 1 - ptukey(qobs, ngroups, (ngroups -
1)*(n - 1))
> ptk
```

Post-Hoc Tests

	p-value
“A” - “B”	0.9600081
“A” - “C”	0.1299235
“B” - “C”	**0.07280894**

Visualizing Results

```
> bargraph.CI(data.mm$Dosage, data.mm$Recall, group=data.mm$Gender, ylim=c(0,25), col="Orange")

> bargraph.CI(data.mm$Dosage, data.mm$Recall, ylim=c(0,25), col="Red")

> bargraph.CI(data.mm$Task, data.mm$Recall, ylim=c(0,25), col="Light Green")
```



Mixed-Model ANOVA using LM

Using Linear Mixed-Effects Model

```
> data.mm=read.table("http://personality-project.org/r/datasets/R.appendix5.data",header=TRUE)

> library(nlme)

> fit.mm.lm = lme(Recall~Task*Dosage, random = ~1|Subject/Task,data=data.mm)

> summary(fit.mm.lm)
```

fit <- lme(y~(W*B), random= ~1|Subject/W, data)

Using Linear Mixed-Effects Model

Difference between Cued and Free Tasks

Difference between dosage A and B

Interaction between Task (Cued and Free) and Dosage (A and B)

```
Linear mixed-effects model fit by REML
 Data: data.mm
      AIC      BIC    logLik
  466.105 489.7297 -224.0525
Random effects:
 Formula: ~1 | Subject
        (Intercept)
StdDev:    4.376845
 Formula: ~1 | Task %in% Subject
        (Intercept) Residual
StdDev:  0.02507055 1.493813
Fixed effects: Recall ~ Task * Dosage
                 Value Std.Error DF   t-value p-value
(Intercept)   14.944444 1.8212277 72  8.205698  0.0000
TaskF         -1.500000 0.4981479 15 -3.011154  0.0088
DosageB       -0.111111 2.5756049 15 -0.043140  0.9662
DosageC        5.000000 2.5756049 15  1.941292  0.0713
TaskF:DosageB -1.166667 0.7044875 15 -1.656050  0.1185
TaskF:DosageC  0.000000 0.7044875 15  0.000000  1.0000
```

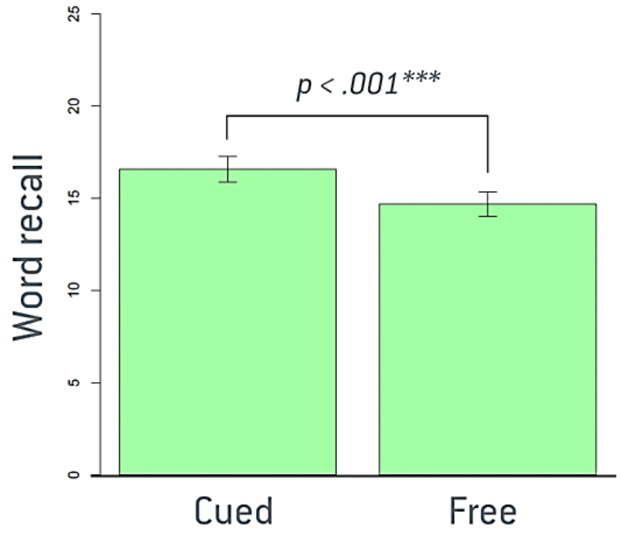
```
Correlation:
              (Intr) TaskF  DosagB DosagC TsF:DB
TaskF         -0.137
DosageB       -0.707  0.097
DosageC       -0.707  0.097  0.500
TaskF:DosageB  0.097 -0.707 -0.137 -0.068
TaskF:DosageC  0.097 -0.707 -0.068 -0.137  0.500
Standardized Within-Group Residuals:
        Min          Q1         Med          Q3 
Max
-3.31140143 -0.46795626  0.05238567  0.56673475 
2.15800723

Number of Observations: 108
Number of Groups:
         Subject Task %in% Subject
              18                36
```

Linear Mixed-Effects Model Post-Hoc Tests

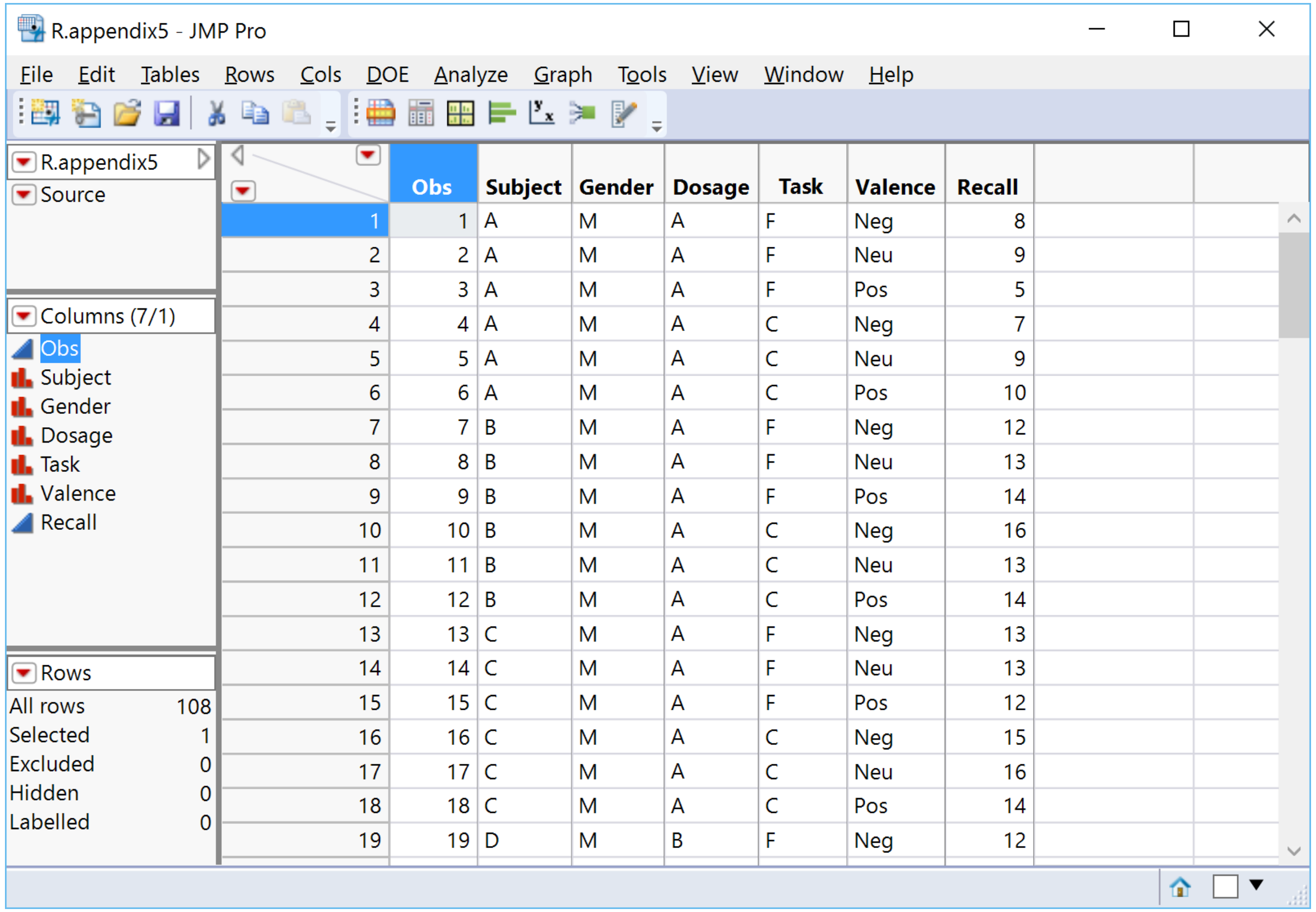
```
> require(multcomp)
> summary(glht(fit.mm.lm, linfct=mcp(Dosage="Tukey")))
Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lme.formula(fixed = Recall ~ Task * Dosage, data = data.mm, random = ~1 |
    Subject/Task)
Linear Hypotheses:
           Estimate Std. Error z value Pr(>|z|)
B - A == 0  -0.1111     2.5756  -0.043    0.999
C - A == 0   5.0000     2.5756   1.941    0.127
C - B == 0   5.1111     2.5756   1.984    0.116
(Adjusted p values reported -- single-step method)
Warning message:
In mcp2matrix(model, linfct = linfct) :
  covariate interactions found -- default contrast might be inappropriate
```

Reporting Results

"We fit a linear mixed-effects model to our data to test how participants' word recall was affected by drug dosage (between-participants factor) and memory task (within-participants factor). The results showed that participants' word recall was better with the cued memory task than with the free memory task, $t(15) = -3.011$, $p < .001$. Pairwise comparisons showed no significant effects across dosage levels over word recall. Also, no significant interactions were identified between dosage level and memory task."

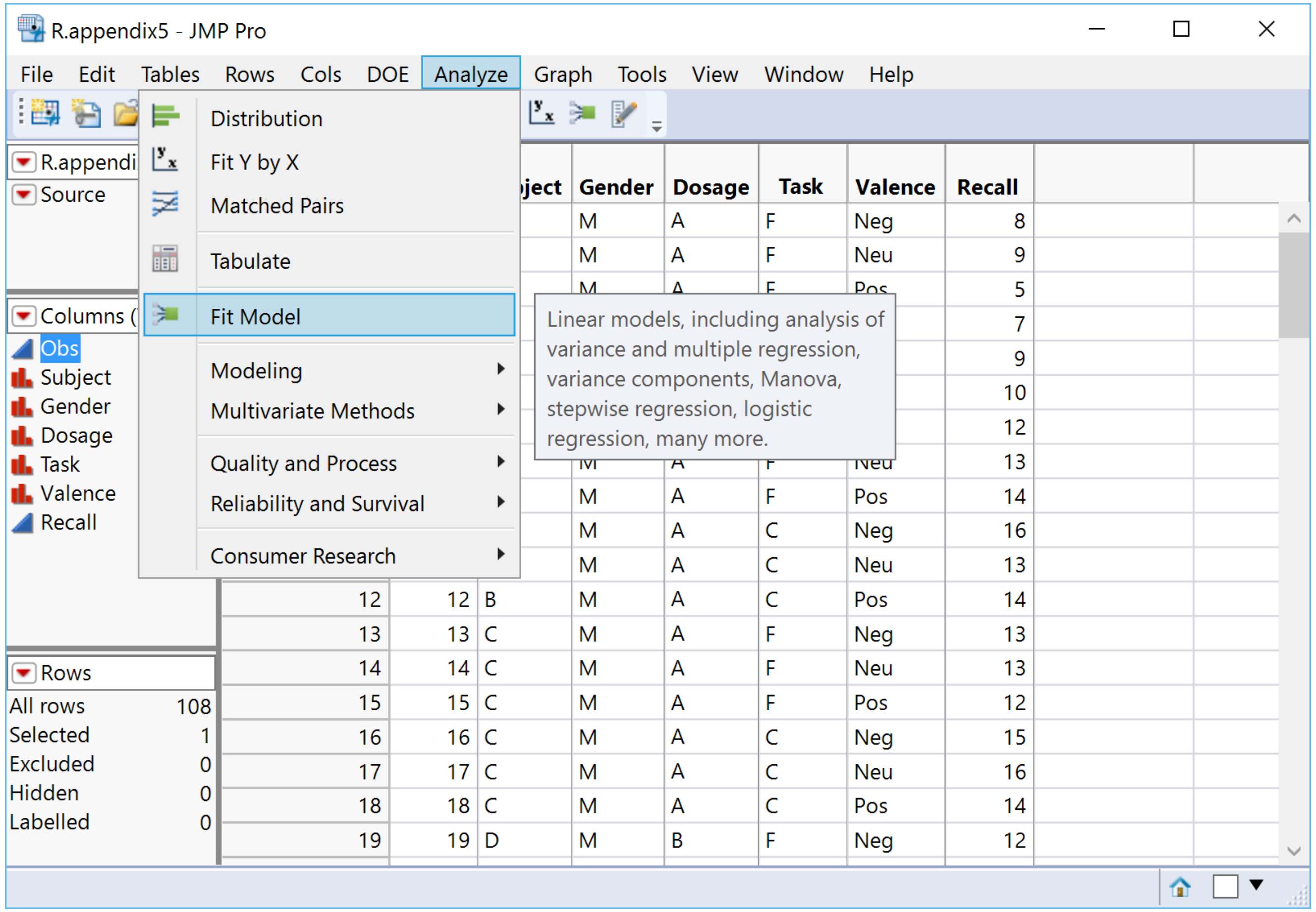


Mixed-Model ANOVA in JMP

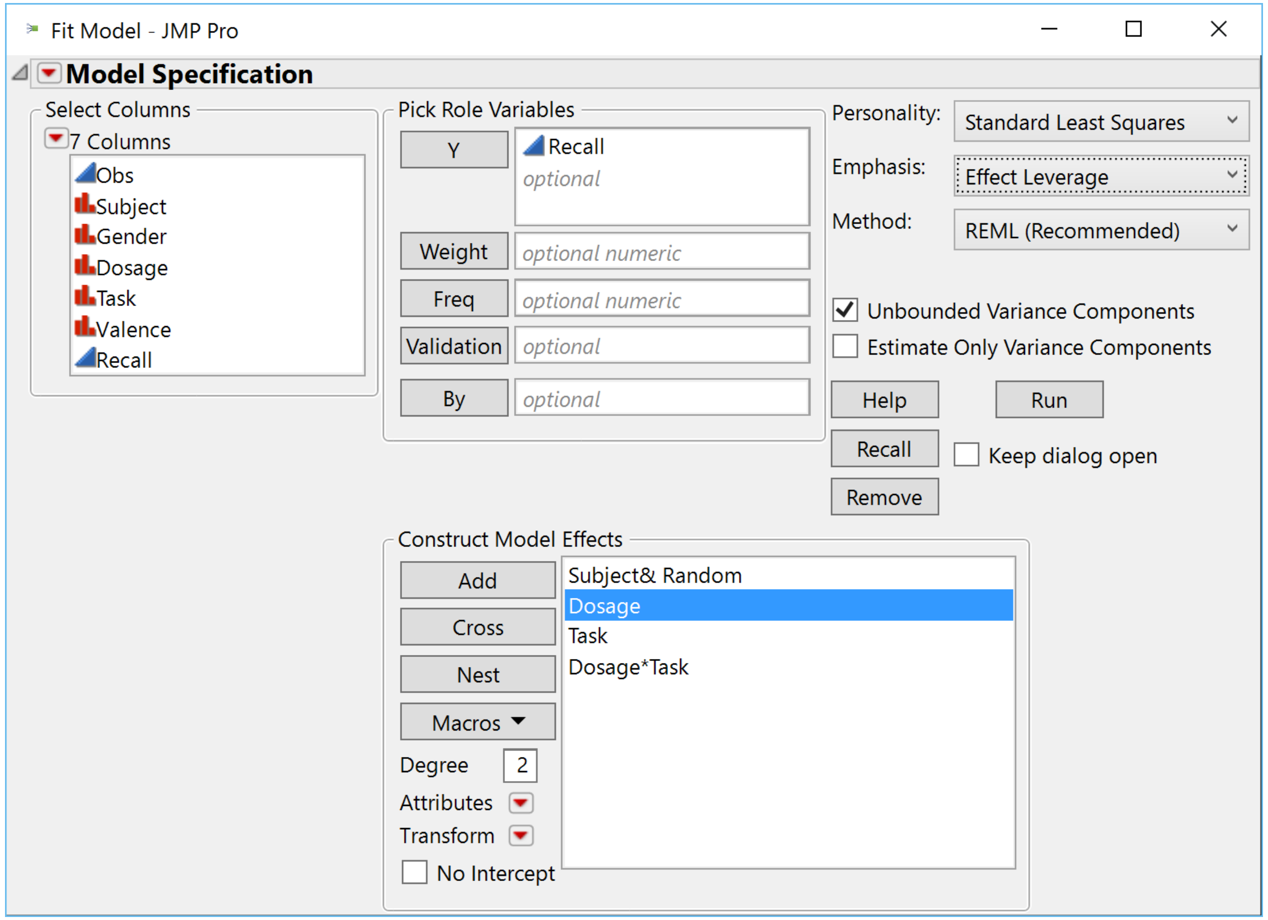


	Obs	Subject	Gender	Dosage	Task	Valence	Recall
1	1	A	M	A	F	Neg	8
2	2	A	M	A	F	Neu	9
3	3	A	M	A	F	Pos	5
4	4	A	M	A	C	Neg	7
5	5	A	M	A	C	Neu	9
6	6	A	M	A	C	Pos	10
7	7	B	M	A	F	Neg	12
8	8	B	M	A	F	Neu	13
9	9	B	M	A	F	Pos	14
10	10	B	M	A	C	Neg	16
11	11	B	M	A	C	Neu	13
12	12	B	M	A	C	Pos	14
13	13	C	M	A	F	Neg	13
14	14	C	M	A	F	Neu	13
15	15	C	M	A	F	Pos	12
16	16	C	M	A	C	Neg	15
17	17	C	M	A	C	Neu	16
18	18	C	M	A	C	Pos	14
19	19	D	M	B	F	Neg	12

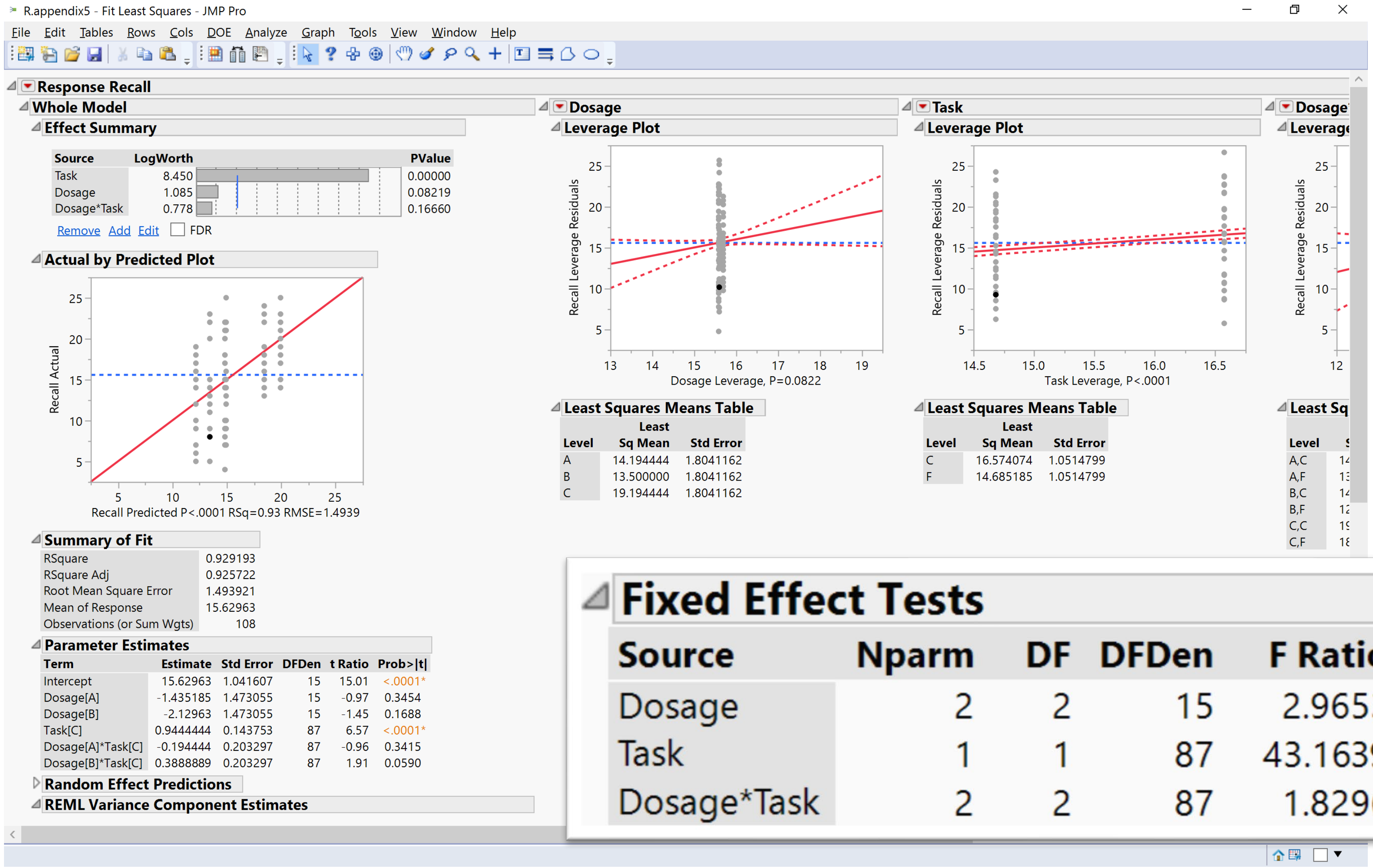
Mixed-Model ANOVA in JMP



Mixed-Model ANOVA in JMP



Mixed-Model ANOVA in JMP

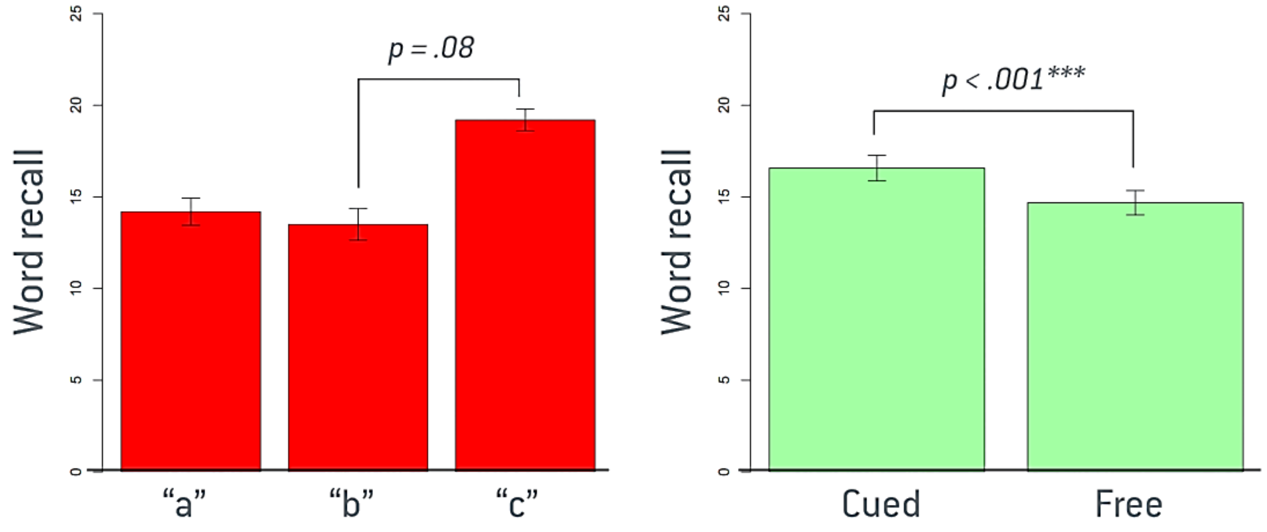


Fixed Effect Tests

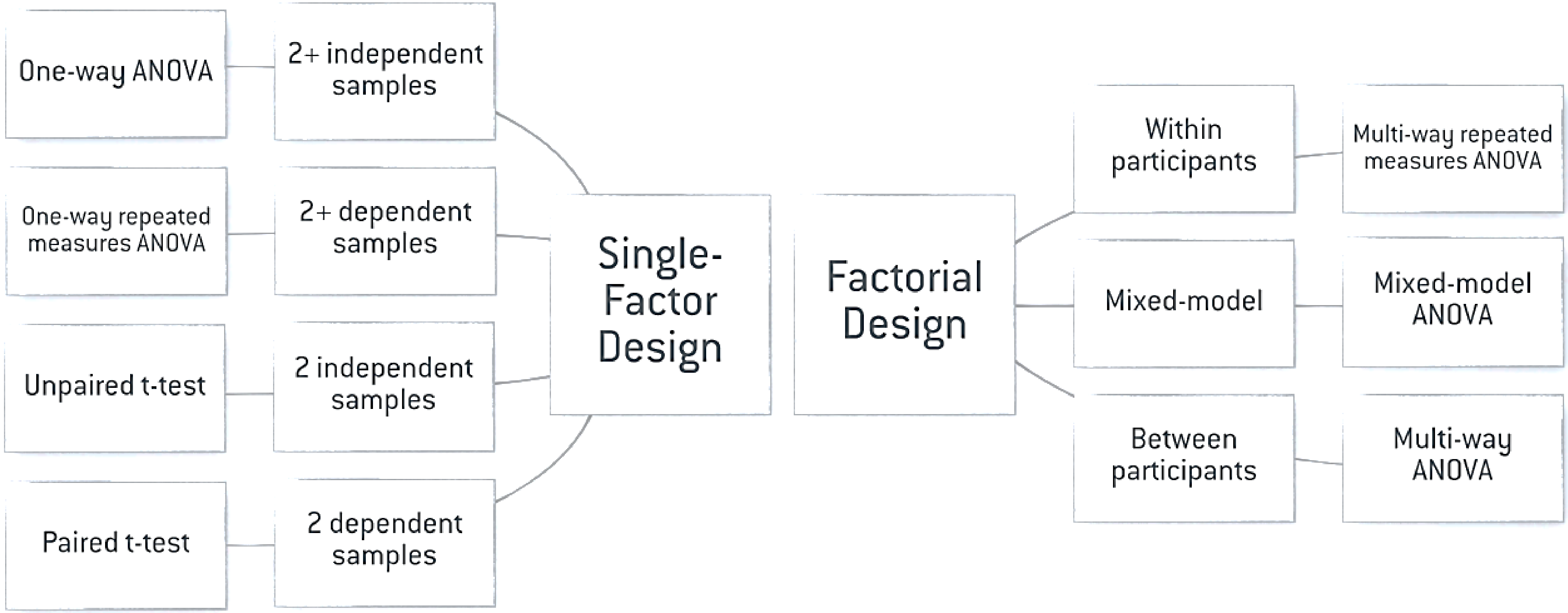
Source	Nparm	DF	DFDen	F Ratio	Prob > F
Dosage	2	2	15	2.9653	0.0822
Task	1	1	87	43.1639	<.0001*
Dosage*Task	2	2	87	1.8296	0.1666

Reporting Results

"We conducted a mixed-model analysis of variance (ANOVA) to test how the memory task (within-participants factor) and drug dosage (between-participants factor) affected participants' recall of the words. The analysis showed that participants had significantly better recall of words when they memorized the words using the cued task than when they did so using the free task, $F(1,15) = 43.13$, $p < .001$. The analysis also showed a marginal difference in word recall among dosage levels, $F(2,15) = 2.97$, $p = .08$. Post-hoc pairwise tests showed that word recall was marginally higher with dosage c than it was with dosage b, $p = .08$. No significant interactions were identified between drug dosage and memory task."



Hypothesis Testing Decision Map



Output Variable

Input Variable

	Nominal	Categorical (>2 categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Nominal	X^2 or Fisher's	X^2	X^2 trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank (a)	Student's t-test
Categorical (>2 categories)	X^2	X^2	Kruskal-Wallis (b)	Kruskal-Wallis (b)	Kruskal-Wallis (b)	**Analysis of variance (c)**
Ordinal	X^2 trend or Mann-Whitney	(e)	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression (d)
Quantitative Discrete	Logistic regression	(e)	(e)	Spearman rank	Spearman rank	Spearman rank or linear regression (d)
Quantitative Non-Normal	Logistic regression	(e)	(e)	(e)	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
Quantitative Normal	Logistic regression	(e)	(e)	(e)	Linear regression (d)	Pearson and linear regression

Questions?

Effect Size

A final note on inferential statistics

Effect Size

- Problems with p-values
  - Null hypothesis is almost never literally true, rejecting it is relatively uninformative
  - Significance dependent on both effect size and sample size
    - Low sample size may lead to Type II errors
    - High sample size can find significance for even trivial effects
- Effect sizes estimate the magnitude of the effect in a manner relatively independent of sample size
  - Tells us how strongly two or more variables are related or how large the difference between groups is
  - Increasingly important to report

Ways of reporting

Eta squared η^2

Partial eta squared $\eta_p{}^2$

Omega squared ω^2

Epsilon squared ε^2

Cohen's *d*

…

Partial eta squared

Recommended way of reporting effect size by APA

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

See: Levine, Timothy R., and Craig R. Hullett. "Eta squared, partial eta squared, and misreporting of effect size in communication research." *Human Communication Research* 28.4 (2002): 612-625.

JMP Example

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

Effect size of **condition**:

$$\eta_p^2 = \frac{95.72917}{95.72917 + 503.83333}$$

$$\eta_p^2 = 0.15967$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	301.97917	43.1399	3.4249
Error	40	503.83333	12.5958	Prob > F
C. Total	47	805.81250		0.0058*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.4375	0.512263	22.33	<.0001*
condition[0-None]	-1.6875	0.887265	-1.90	0.0644
condition[1-Mal]	-1.104167	0.887265	-1.24	0.2206
condition[2-Ad]	1.3125	0.887265	1.48	0.1469
gender[Female]	-0.895833	0.512263	-1.75	0.0880
gender[Female]*condition[0-None]	0.6458333	0.887265	0.73	0.4709
gender[Female]*condition[1-Mal]	1.5625	0.887265	1.76	0.0859
gender[Female]*condition[2-Ad]	0.9791667	0.887265	1.10	0.2764

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
condition	3	3	95.72917	2.5334	0.0705
gender	1	1	38.52083	3.0582	0.0880
gender*condition	3	3	167.72917	4.4387	0.0088*

JMP Example

$$\eta_p^2 = 0.15967$$

Reporting:

A marginal main effect of type of review on learning, F(3,40) = 2.53, p = .071, η_p^2 = .160.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	301.97917	43.1399	3.4249
Error	40	503.83333	12.5958	**Prob > F**
C. Total	47	805.81250		0.0058*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.4375	0.512263	22.33	<.0001*
condition[0-None]	-1.6875	0.887265	-1.90	0.0644
condition[1-Mal]	-1.104167	0.887265	-1.24	0.2206
condition[2-Ad]	1.3125	0.887265	1.48	0.1469
gender[Female]	-0.895833	0.512263	-1.75	0.0880
gender[Female]*condition[0-None]	0.6458333	0.887265	0.73	0.4709
gender[Female]*condition[1-Mal]	1.5625	0.887265	1.76	0.0859
gender[Female]*condition[2-Ad]	0.9791667	0.887265	1.10	0.2764

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
condition	3	3	95.72917	2.5334	0.0705
gender	1	1	38.52083	3.0582	0.0880
gender*condition	3	3	167.72917	4.4387	0.0088*

Szafir, Daniel, and Bilge Mutlu. "ARTFul: adaptive review technology for flipped learning." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013.

Next

Individual Assignment 4: Statistical Analysis

- Due midnight Sunday 4/28
- See guidelines on Moodle

Reading #9:

- Due Monday 4/29

Group Project

- Final presentations Wednesday 5/1
- Final write-up due Wednesday 5/6 at 10:00pm
- (you can collect additional data between presentation and final write-up)

University of Colorado
Boulder

THANKS!

Professor **Dan Szafir**

Computer Science & ATLAS Institute
University of Colorado Boulder