Crossed Random Effects for Other (Not Longitudinal) Repeated Measures Designs

- Topics:
  - **ANOVA for repeated measures**
  - MLM for repeated measures

Analytic Toolbox of the Experimental Psychologist

- Our friend, analysis of variance (ANOVA)
  - Between-group (*aka* between-subject, independent IV)
  - Within-group (*aka* within-subject, dependent, repeated measures IV)
  - Split-plot (*aka* mixed design of between- and within-group IVs)

- Expandable to include:
  - multiple IVs (factorial ANOVA)
  - main effects of continuous covariates (ANCOVA)
  - multiple outcomes (MANOVA/MANCOVA)

ANOVA works well when…

- Experimental stimuli are ***controlled*** and ***exchangeable***
  - Controlled → Constructed, not sampled from a population
  - Exchangeable → Stimuli vary only in dimensions of interest
  - …What to do with non-exchangeable stimuli (e.g., words, scenes)?

- Experimental manipulations create ***discrete conditions***
  - e.g., set size of 3 vs. 6 vs. 9 items
  - e.g., response compatible vs. incompatible distractors
  - …What to do with *continuous* item predictors (e.g., time, salience)?

- One has ***complete data***
  - e.g., if outcome is RT and accuracy is near ceiling
  - e.g., if responses are missing for no systematic reason
  - …What if data are not missing completely at random (e.g., inaccuracy)?

Motivating Example: Psycholinguistic Study Designs

- Word Recognition Tasks (e.g., Lexical Decision)
  - Word lists are constructed based on targeted dimensions while controlling for other relevant dimensions
  - Outcome = RT to decide if the stimulus is a word or non-word (accuracy is usually near ceiling)

- Tests of effects of experimental treatment are typically conducted with the person as the unit of analysis…
  - Average the responses over words within conditions
    - Contentious fights with reviewers about adequacy of experimental control when using real words as stimuli
    - Long history of debate as to how words as experimental stimuli should be analyzed… F_1 ANOVA or F_2 ANOVA (or both)?
    - F_1 only creates a "Language-as-Fixed-Effects Fallacy" (Clark, 1973)

ANOVAs on Summary Data

Original Data per Subject

	B1	B2
A1	Trial 001 Trial 002 .... Trial 100	Trial 101 Trial102 .... Trial 200
A2	Trial 201 Trial 202 .... Trial 300	Trial 301 Trial302 .... Trial 400

Subject Summary Data

	B1	B2
A1	Mean (A1, B1)	Mean (A1, B2)
A2	Mean (A2, B1)	Mean (A2, B2)

"F_1" Repeated Measures ANOVA on *N* subjects:

$$\mathrm{RT}_{cs} = \gamma_0 + \gamma_1 \mathrm{A}_c + \gamma_2 \mathrm{B}_c + \gamma_3 \mathrm{A}_c \mathrm{B}_c + \mathbf{U}_{0s} + \mathrm{e}_{cs}$$

"F_2" Between-Groups ANOVA on *T* trials:

$$\mathrm{RT}_t = \gamma_0 + \gamma_1 \mathrm{A}_t + \gamma_2 \mathrm{B}_t + \gamma_3 \mathrm{A}_t \mathrm{B}_t + \mathrm{e}_t$$

Trial Summary Data

	B1
A1, B1	Trial 001 = Mean(Subject 1, Subject 2,... Subject *N*) Trial 002 = Mean(Subject 1, Subject 2,... Subject *N*) .... Trial 100
A1, B2	Trial 101 = Mean(Subject 1, Subject 2,... Subject *N*) Trial 102 = Mean(Subject 1, Subject 2,... Subject *N*) .... Trial 200
A2, B1	Trial 201 = Mean(Subject 1, Subject 2,... Subject *N*) Trial 202 = Mean(Subject 1, Subject 2,... Subject *N*) .... Trial 300
A2, B2	Trial 301 = Mean(Subject 1, Subject 2,... Subject *N*) Trial 302 = Mean(Subject 1, Subject 2,... Subject *N*) .... Trial 400

Choosing Amongst ANOVA Models

- F_1 RM ANOVA on **subject** summary data:
  - Assumes trials are fixed—within-condition ***trial*** variability is gone
- F_2 ANOVA on **trial** summary data:
  - Assumes persons are fixed—within-trial ***subject*** variability is gone

- Proposed ANOVA-based resolutions:
  - **F′** → quasi-F test that treats both trials and subjects as random (Clark, 1973), but requires complete data (least squares)
  - **Min F′** → lower-bound of F′ derived from F1 and F2 results, which does not require complete data, but is (too) conservative
  - **F_1 x F_2 criterion** → effects are only "real" if they are significant in **both F_1 and F_2 models** (aka, death knell for psycholinguists)

  - But neither model is complete (two wrongs don't make a right)…

Sources of Variance (Clark, 1973)

t = #conditions, i = #items, s = #subjects

Label		DF	Expected Mean Square
T	Treatments (t)	t−1	$\sigma^2_e + \sigma^2_{SxI} + i\sigma^2_{TxS} +$ ___ $+ s\sigma^2_I + is\sigma^2_T$
I w T	Items (i) within Treatments	t(i−1)	$\sigma^2_e + \sigma^2_{SxI} +$ ___ $+$ ___ $+ s\sigma^2_I +$ ___
S	Subjects (s)	s−1	$\sigma^2_e + \sigma^2_{SxI} +$ ___ $+ t\sigma^2_S +$ ___ $+$ ___
T x S	Treatments by Subjects	(t−1)(s−1)	$\sigma^2_e + \sigma^2_{SxI} + i\sigma^2_{TxS} +$ ___ $+$ ___ $+$ ___
S x I w T	Subjects by Items within Treatments	t(i−1)(s−1)	$\sigma^2_e + \sigma^2_{SxI} +$ ___ $+$ ___ $+$ ___ $+$ ___

Effect of Treatment via F_1 ANOVA

T numerator should differ from TxS denominator by 1 term

Label		DF	Expected Mean Square
T	**Treatments (t)**	**t−1**	$\sigma_e^2 + \sigma_{SxI}^2 + i\sigma_{TxS}^2 +$ ___ $+ \boxed{s\sigma_I^2} + \boxed{is\sigma_T^2}$
I w T	Items (i) within Treatments	t(i−1)	$\sigma_e^2 + \sigma_{SxI}^2 +$ ___ $+$ ___ $+ s\sigma_I^2 +$ ___
S	Subjects (s)	s−1	$\sigma_e^2 + \sigma_{SxI}^2 +$ ___ $+ t\sigma_S^2 +$ ___ $+$ ___
T x S	**Treatments by Subjects**	**(t−1)(s−1)**	$\sigma_e^2 + \sigma_{SxI}^2 + i\sigma_{TxS}^2 +$ ___ $+$ ___ $+$ ___
S x I w T	Subjects by Items within Treatments	t(i−1)(s−1)	$\sigma_e^2 + \sigma_{SxI}^2 +$ ___ $+$ ___ $+$ ___ $+$ ___

Effect of Treatment via F_2 ANOVA

T numerator should differ from IxT denominator by 1 term

Label		DF	Expected Mean Square
T	Treatments (t)	$t-1$	$\sigma_e^2 + \sigma_{SxI}^2 + \boxed{i\sigma_{TxS}^2} +$ ___ $+ s\sigma_I^2 + \boxed{is\sigma_T^2}$
I w T	Items (i) within Treatments	$t(i-1)$	$\sigma_e^2 + \sigma_{SxI}^2 +$ ___ $+$ ___ $+ s\sigma_I^2 +$ ___
S	Subjects (s)	$s-1$	$\sigma_e^2 + \sigma_{SxI}^2 +$ ___ $+ t\sigma_S^2 +$ ___ $+$ ___
T x S	Treatments by Subjects	$(t-1)(s-1)$	$\sigma_e^2 + \sigma_{SxI}^2 + i\sigma_{TxS}^2 +$ ___ $+$ ___ $+$ ___
S x I w T	Subjects by Items within Treatments	$t(i-1)(s-1)$	$\sigma_e^2 + \sigma_{SxI}^2 +$ ___ $+$ ___ $+$ ___ $+$ ___

Simultaneous Quasi-F Ratio (F')

- F′ was proposed by Clark (1973) as a quasi-F test that treats both items and subjects as random factors

$$F'\left(df_{num}, df_{den}\right) = \frac{MS_T + MS_{SxI}}{MS_{TxS} + MS_I}$$

$$\text{where } df_{num} = \frac{\left(MS_T + MS_{SxI}\right)^2}{\frac{MS_T}{df_T} + \frac{MS_{SxI}}{df_{SxI}}} \text{ and } df_{den} = \frac{\left(MS_{TxS} + MS_I\right)^2}{\frac{MS_{TxS}}{df_{TxS}} + \frac{MS_I}{df_I}}$$

$$F'\left(df_{num}, df_{den}\right) = \frac{\left(2*\sigma_e^2\right) + \left(2*\sigma_{SxI}^2\right) + \left(\#I*\sigma_{TxS}^2\right) + \left(\#S*\sigma_I^2\right) + \left(\#I*\#S*\sigma_T^2\right)}{\left(2*\sigma_e^2\right) + \left(2*\sigma_{SxI}^2\right) + \left(\#I*\sigma_{TxS}^2\right) + \left(\#S*\sigma_I^2\right)}$$

- Numerator then exceeds the denominator by exactly the treatment variance as desired… except it requires complete data given that it relies on least squares
  - Not feasible in most real-world experiments

Minimum of Quasi-F Ratio (Min F′)

- Min F′ was developed to be used from F_1 and F_2 results:

$$\min F'\left(df_{num}, df_{den}\right) = \frac{MS_T}{MS_{TxS} + MS_I} = \frac{F_1 * F_2}{F_1 + F_2}$$

- But given that Min F′ is overly conservative, having to show significance by both models is often required instead:
  - the F_1 by F_2 criterion… but two wrongs don't make a right

- Wouldn't it be nice if we had some way to treat subjects and items as the random effects they actually are???
  - And to assess the extent to which items are actually exchangeable?
  - And that all the extraneous item variables were adequately controlled?
  - **Multilevel models to the rescue! … maybe?**

Crossed Random Effects for Other (Not Longitudinal) Repeated Measures Designs

- Topics:
  - ANOVA for repeated measures
  - **MLM for repeated measures**

Multilevel Models to the Rescue?

Original Data per Person

	B1	B2
A1	Trial 001 Trial 002 .... Trial 100	Trial 101 Trial102 .... Trial 200
A2	Trial 201 Trial 202 .... Trial 300	Trial 301 Trial302 .... Trial 400

Pros:

- Use all original data, not summaries
- Responses can be missing at random
- Can include continuous trial predictors

Cons:

- **Is still wrong**

Level 1: $y_{ts} = \beta_{0s} + \beta_{1s}A_{ts} + \beta_{2s}B_{ts} + \beta_{3s}A_{ts}B_{ts} + e_{ts}$

Level 2: $\beta_{0s} = \gamma_{00} + U_{0s}$

$\beta_{1s} = \gamma_{10}$

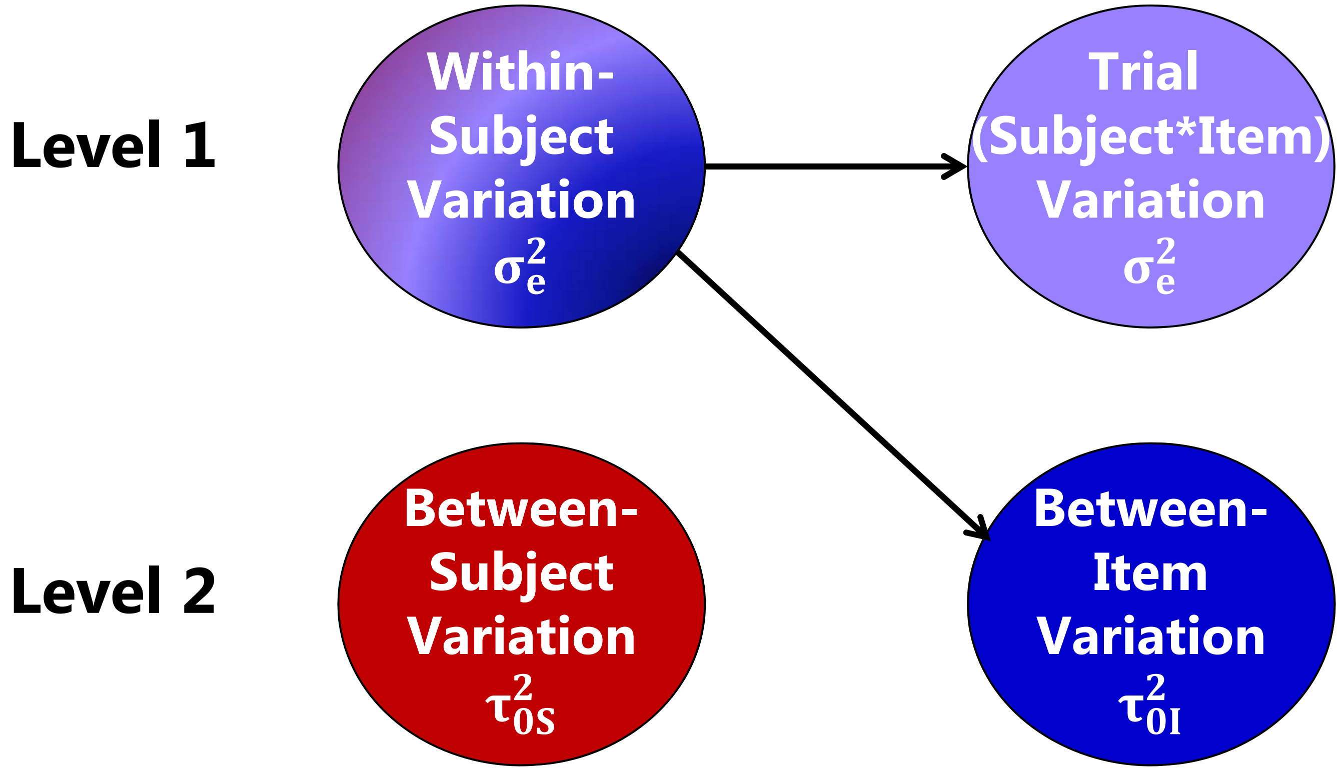
$\beta_{2s} = \gamma_{20}$

$\beta_{3s} = \gamma_{30}$

Level 1 = Within-Subject Variation (Across Trials)

Level 2 = Between-Subject Variation

Multilevel Models to the Rescue?



Empty Means, Crossed Random Effects Models

- **Residual-only model:**
  - $RT_{tis} = \gamma_{000} + e_{tis}$
  - Assumes no effects (dependency) of subjects or items

- **Random subjects model:**
  - $RT_{tis} = \gamma_{000} + \mathbf{U_{00s}} + e_{tis}$
  - Models systematic mean differences **between subjects**

- **Random subjects and items model:**
  - $RT_{tis} = \gamma_{000} + U_{00s} + \mathbf{U_{0i0}} + e_{tis}$
  - Also models systematic mean differences **between items**

A Better Way of (Multilevel) Life

Between-Subject Variation L2 τ^2_{0S}

Between-Item Variation L2 τ^2_{0I}

Trial (Subject*Item) Variation σ^2_e

Random effects over **subjects** of **item** or **trial** predictors can also be tested and predicted.

- **Multilevel Model with *Crossed* Random Effects:**

$$\mathrm{RT}_{tis} = \gamma_{000} + \gamma_{010}\mathrm{A}_i + \gamma_{020}\mathrm{B}_i + \gamma_{030}\mathrm{A}_i\mathrm{B}_i + \mathbf{U}_{00s} + \mathbf{U}_{0i0} + \mathbf{e}_{tis}$$

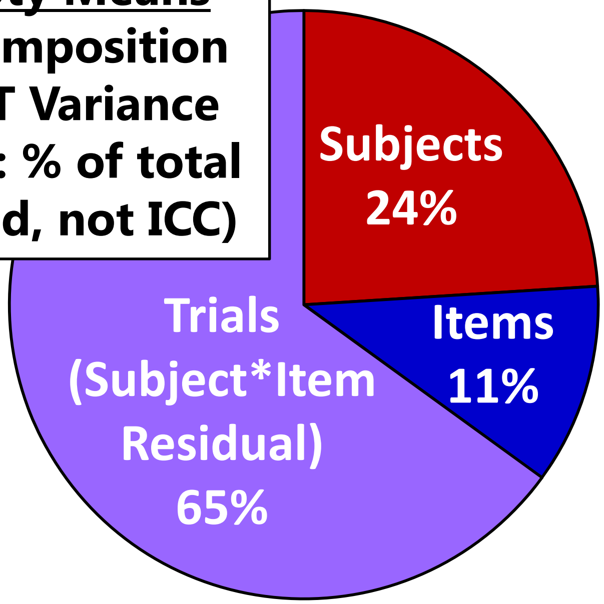
t trial
i item
s subject

- Both **subjects** and **items** as random effects:
  - Subject predictors explain between-subject mean variation: τ^2_{0S}
  - Item predictors explain between-item mean variation: τ^2_{0I}
  - Trial predictors explain trial-specific residual variation: σ^2_e

Example Psycholinguistic Study
(Locker, Hoffman, & Bovaird, 2007)

- Crossed design: 38 subjects by 39 items (words or nonwords)
- Lexical decision task: RT to decide if word or nonword
- 2 word-specific predictors of interest:
  - A: Low/High Phonological Neighborhood Frequency
  - B: Small/Large Semantic Neighborhood Size

Empty Means Decomposition of RT Variance (note: % of total is used, not ICC)



Model and Results

$$RT_{tis} = \gamma_{000} + \gamma_{010}A_i + \gamma_{020}B_i + \gamma_{030}A_iB_i + U_{00s} + U_{0i0} + e_{tis}$$

Pseudo-R^2:

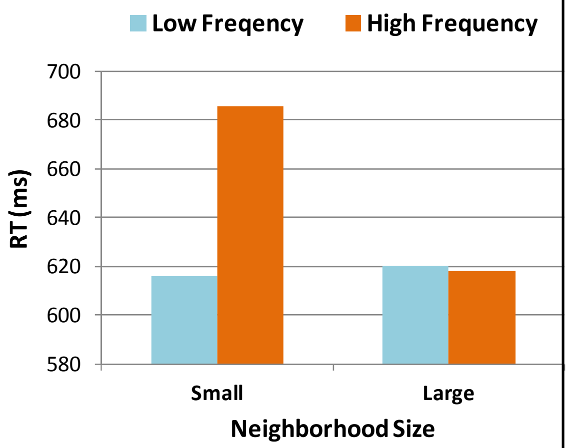
Residual ≈ 0%

Subjects ≈ 0%

Items ≈ 30%*

Total R^2 ≈ 3.3%

***Significant item variability remained**



Tests of Fixed Effects by Model

	A: Frequency Marginal Main Effect	B: Size Marginal Main Effect	A*B: Interaction of Frequency by Size
F_1 Subjects ANOVA	$F\,(1,37) = 16.1$ $p = .0003$	$F\,(1,37) = 14.9$ $p = .0004$	$F\,(1,37) = 38.2$ $p < .0001$
F_2 Words ANOVA	$F\,(1,35) = 5.3$ $p = .0278$	$F\,(1,35) = 4.5$ $p = .0415$	$F\,(1,35) = 5.7$ $p = .0225$
F′ min (via ANOVA)	$F\,(1,56) = 4.0$ $p = .0530$	$F\,(1,55) = 3.5$ $p = .0710$	$F\,(1,45) = 5.0$ $p = .0310$
Crossed MLM (via REML)	$F\,(1,32) = 5.4$ $p = .0272$	$F\,(1,32) = 4.6$ $p = .0393$	$F\,(1,32) = 6.0$ $p = .0199$

Tests of Fixed Effects by Model

	A: Frequency Marginal Main Effect	B: Size Marginal Main Effect	A*B: Interaction of Frequency by Size
F_1 Subjects ANOVA	$F\ (1,37) = 16.1$ $p = .0003$	$F\ (1,37) = 14.9$ $p = .0004$	$F\ (1,37) = 38.2$ $p < .0001$
F_2 Words ANOVA	$F\ (1,35) = 5.3$ $p = .0278$	$F\ (1,35) = 4.5$ $p = .0415$	$F\ (1,35) = 5.7$ $p = .0225$
F′ min (via ANOVA)	$F\ (1,56) = 4.0$ $p = .0530$	$F\ (1,55) = 3.5$ $p = .0710$	$F\ (1,45) = 5.0$ $p = .0310$
Crossed MLM (via REML)	$F\ (1,32) = 5.4$ $p = .0272$	$F\ (1,32) = 4.6$ $p = .0393$	$F\ (1,32) = 6.0$ $p = .0199$

Tests of Fixed Effects by Model

	A: Frequency Marginal Main Effect	B: Size Marginal Main Effect	A*B: Interaction of Frequency by Size
F_1 Subjects ANOVA	$F(1,37) = 16.1$ $p = .0003$	$F(1,37) = 14.9$ $p = .0004$	$F(1,37) = 38.2$ $p < .0001$
F_2 Words ANOVA	$F(1,35) = 5.3$ $p = .0278$	$F(1,35) = 4.5$ $p = .0415$	$F(1,35) = 5.7$ $p = .0225$
F′ min (via ANOVA)	$F(1,56) = 4.0$ $p = .0530$	$F(1,55) = 3.5$ $p = .0710$	$F(1,45) = 5.0$ $p = .0310$
Crossed MLM (via REML)	$F(1,32) = 5.4$ $p = .0272$	$F(1,32) = 4.6$ $p = .0393$	$F(1,32) = 6.0$ $p = .0199$

Simulation: Type 1 Error Rates

Condition			Models					
Item Variance	Subject Variance		**1: Both Random Effects**	2: Random Subjects Only	3: Random Items Only	4: No Random Effects	5: F1 Subjects ANOVA	6: F2 Item ANOVA
Item Effect:								
2	2		**0.03**	0.09	0.03	0.09	0.09	0.03
2	10		**0.05**	0.14	0.05	0.12	0.15	0.05
10	2		**0.04**	0.32	0.04	0.31	0.32	0.04
10	10		**0.05**	0.31	0.05	0.29	0.33	0.05
Subject Effect:								
2	2		**0.04**	0.04	0.12	0.11	0.04	0.12
2	10		**0.05**	0.05	0.34	0.34	0.05	0.36
10	2		**0.04**	0.03	0.12	0.09	0.03	0.12
10	10		**0.06**	0.06	0.34	0.31	0.05	0.37

Model Items as Fixed → Wrong Item Effect

Condition			Models					
Item Variance	Subject Variance		1: Both Random Effects	**2: Random Subjects Only**	3: Random Items Only	4: No Random Effects	**5: F1 Subjects ANOVA**	6: F2 Item ANOVA
Item Effect:								
2	2		0.03	**0.09**	0.03	0.09	**0.09**	0.03
2	10		0.05	**0.14**	0.05	0.12	**0.15**	0.05
10	2		0.04	**0.32**	0.04	0.31	**0.32**	0.04
10	10		0.05	**0.31**	0.05	0.29	**0.33**	0.05
Subject Effect:								
2	2		0.04	0.04	0.12	0.11	0.04	0.12
2	10		0.05	0.05	0.34	0.34	0.05	0.36
10	2		0.04	0.03	0.12	0.09	0.03	0.12
10	10		0.06	0.06	0.34	0.31	0.05	0.37

Model Subjects as Fixed → Wrong Subject Effect

Condition			Models					
Item Variance	Subject Variance		1: Both Random Effects	2: Random Subjects Only	**3: Random Items Only**	4: No Random Effects	5: F1 Subjects ANOVA	**6: F2 Item ANOVA**
Item Effect:								
2	2		0.03	0.09	0.03	0.09	0.09	0.03
2	10		0.05	0.14	0.05	0.12	0.15	0.05
10	2		0.04	0.32	0.04	0.31	0.32	0.04
10	10		0.05	0.31	0.05	0.29	0.33	0.05
Subject Effect:								
2	2		0.04	0.04	**0.12**	0.11	0.04	**0.12**
2	10		0.05	0.05	**0.34**	0.34	0.05	**0.36**
10	2		0.04	0.03	**0.12**	0.09	0.03	**0.12**
10	10		0.06	0.06	**0.34**	0.31	0.05	**0.37**

Conclusions

- An ANOVA model may be less than ideal when:
  - Stimuli are not completely controlled or exchangeable
  - Experimental conditions are not strictly discrete
  - Missing data may result in bias, a loss of power, or both

- ANOVA is a special case of a more general family of multilevel models (with nested or crossed effects as needed) that can offer additional flexibility:
  - Useful in addressing statistical problems →
    - Dependency, heterogeneity of variance, unbalanced or missing data
    - Examine predictor effects pertaining to each source of variation more accurately given that all variation is properly represented in the model
  - Useful in addressing substantive hypotheses →
    - Examining individual differences in effects of experimental manipulations