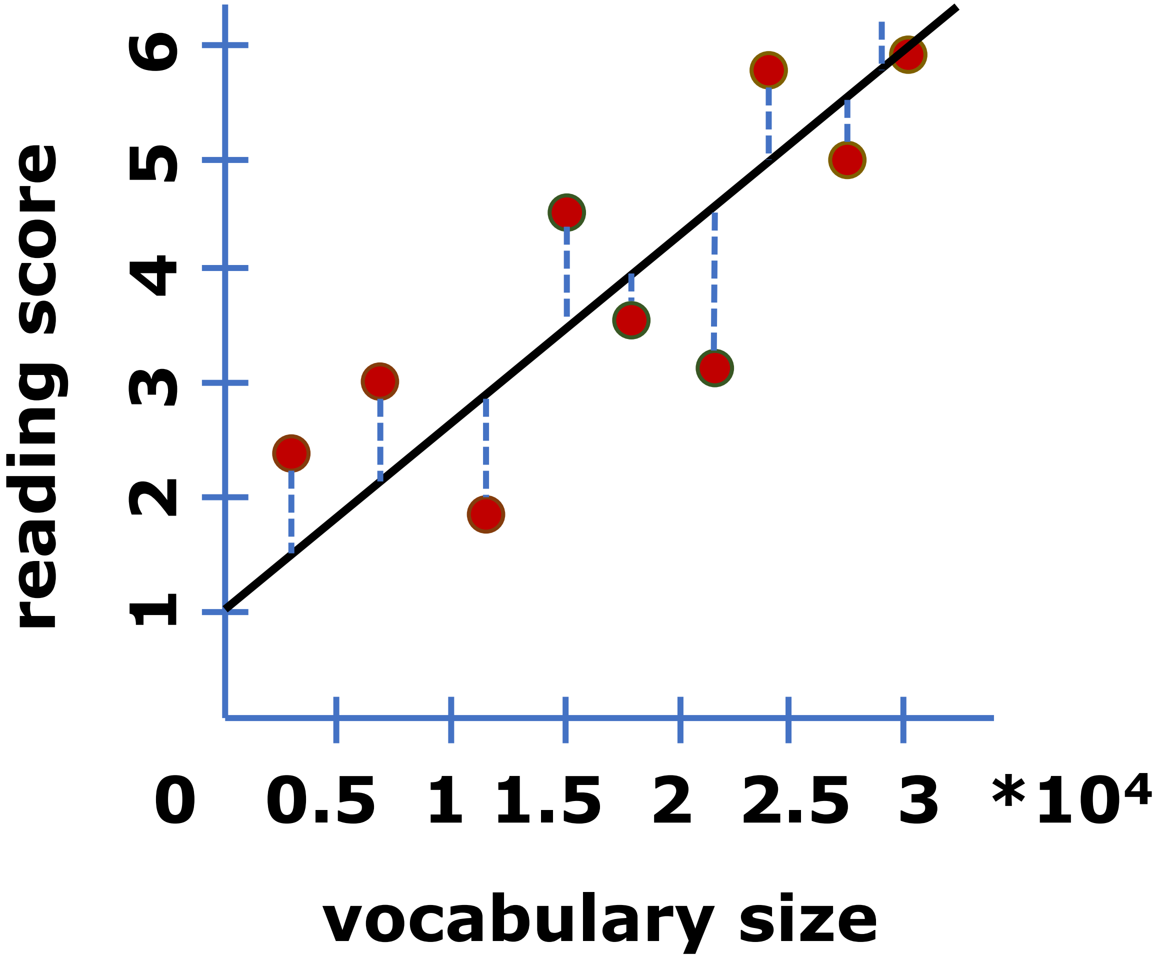
Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

Fundamentals of Statistics for Language Sciences LT2206

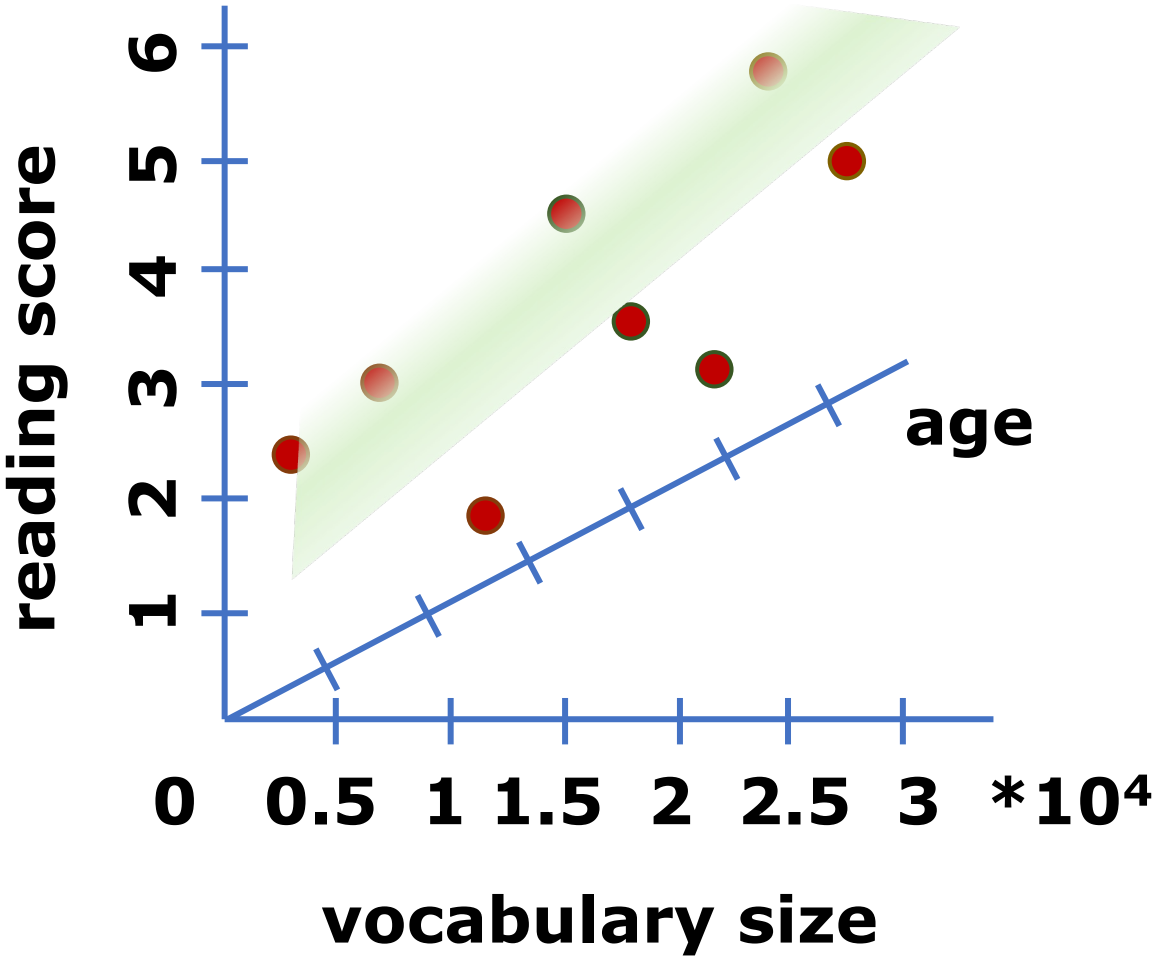
Jixing Li

Lecture 12: Linear Mixed-effects Model

Simple linear regression & multiple regression



$$y = b_1 x + b_0$$



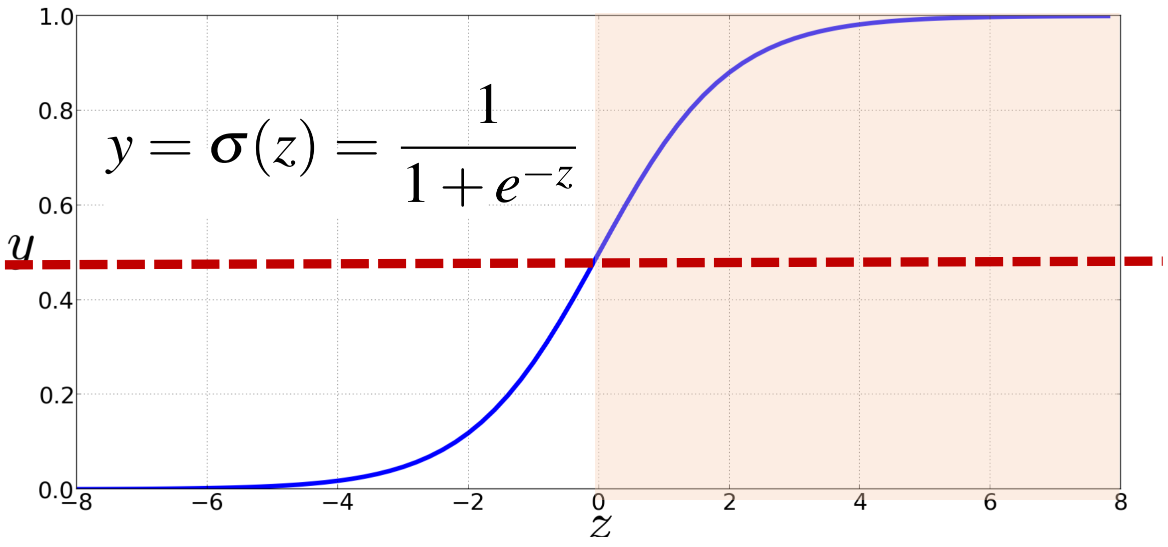
$$y = b_1 x_1 + b_2 x_2 + b_0$$

Logistic regression

When the dependent variable is binary
→ a classification task: a name is male or female

$$y = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\exp(-z)}$$

→ **the sigmoid function**



→ **decision boundary**

$$\hat{y} = \begin{cases} 1 & \text{if } w\cdot x+b > 0 \\ 0 & \text{if } w\cdot x+b \le 0 \end{cases}$$

Example

[卓，琳, Cheuk, Lam, LLA]

$x = [0.5, 0.7, 0.5, 0.6, 0.8]$

$w = [0.1, 0.8, -0.1, 0.2, 0.7]$

$z = w \cdot x + b$

$= w_1{*}x_1 + w_2{*}x_2 + w_3{*}x_3 + w_4{*}x_4 + w_5{*}x_5 + b$

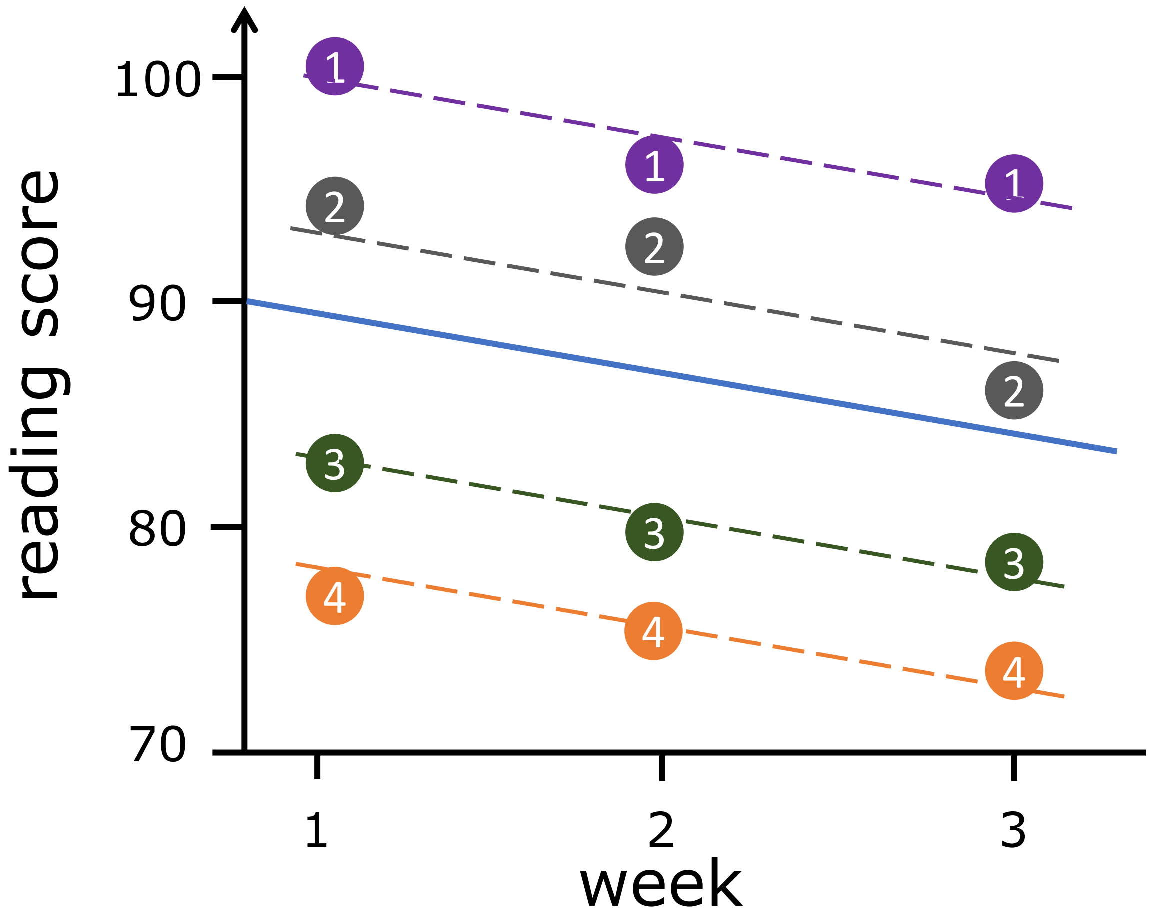
$= 0.05 + 0.56 + (-0.05) + 0.12 + 0.56 + 0.3$

$= 1.54$

$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-1.54}} = 0.82 > 0.5 \rightarrow \text{female}$

Linear mixed-effects model (LMM)

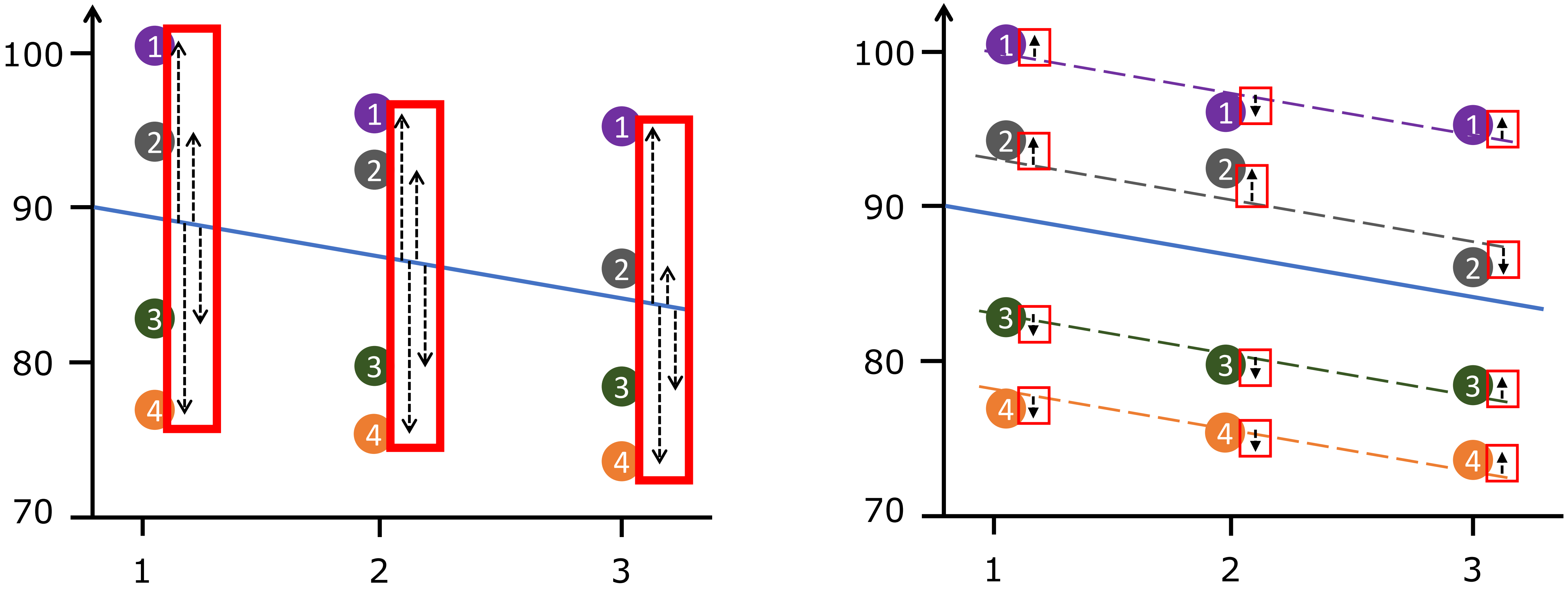
When to use: Studies that obtain multiple measurements over time (longitudinal, time-series) or multiple trials per participant (within subjects)



Fixed effects: variables you are interested in, population-level variables

Random effects: uncontrollable variability, subject-level variance

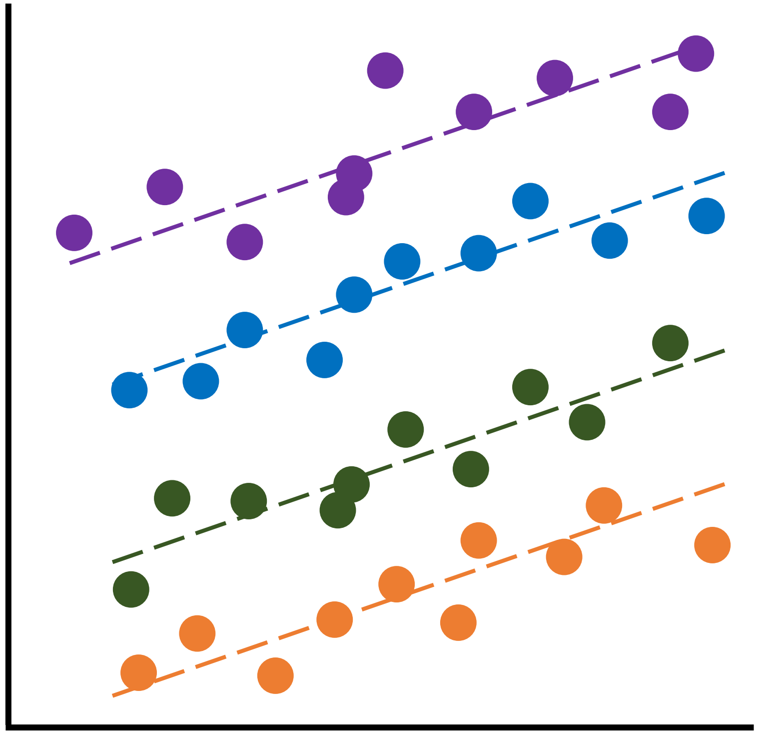
LMM vs. LM



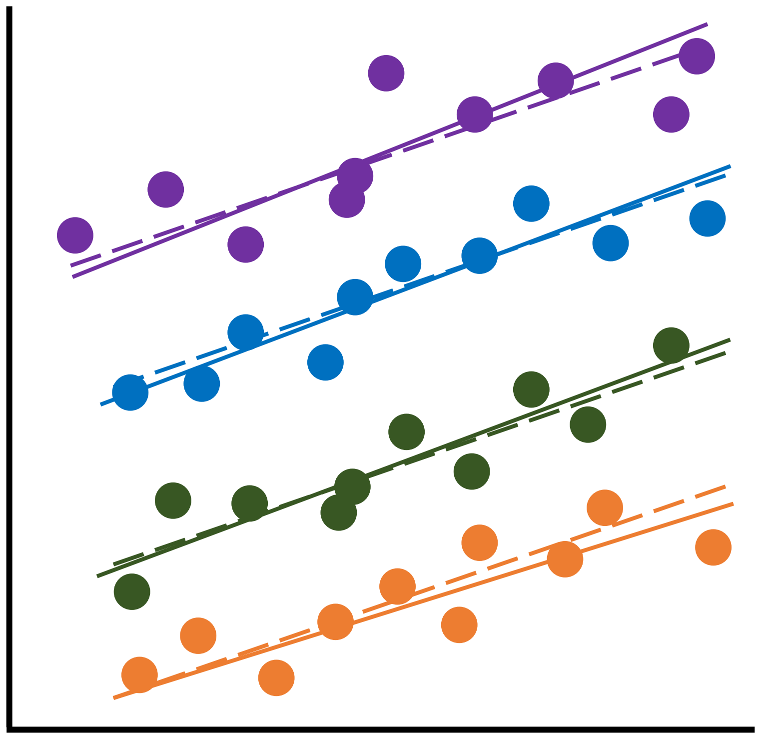
Accounting for individual differences reduces Sum of Squares Total (SST)

Random slope and intercept

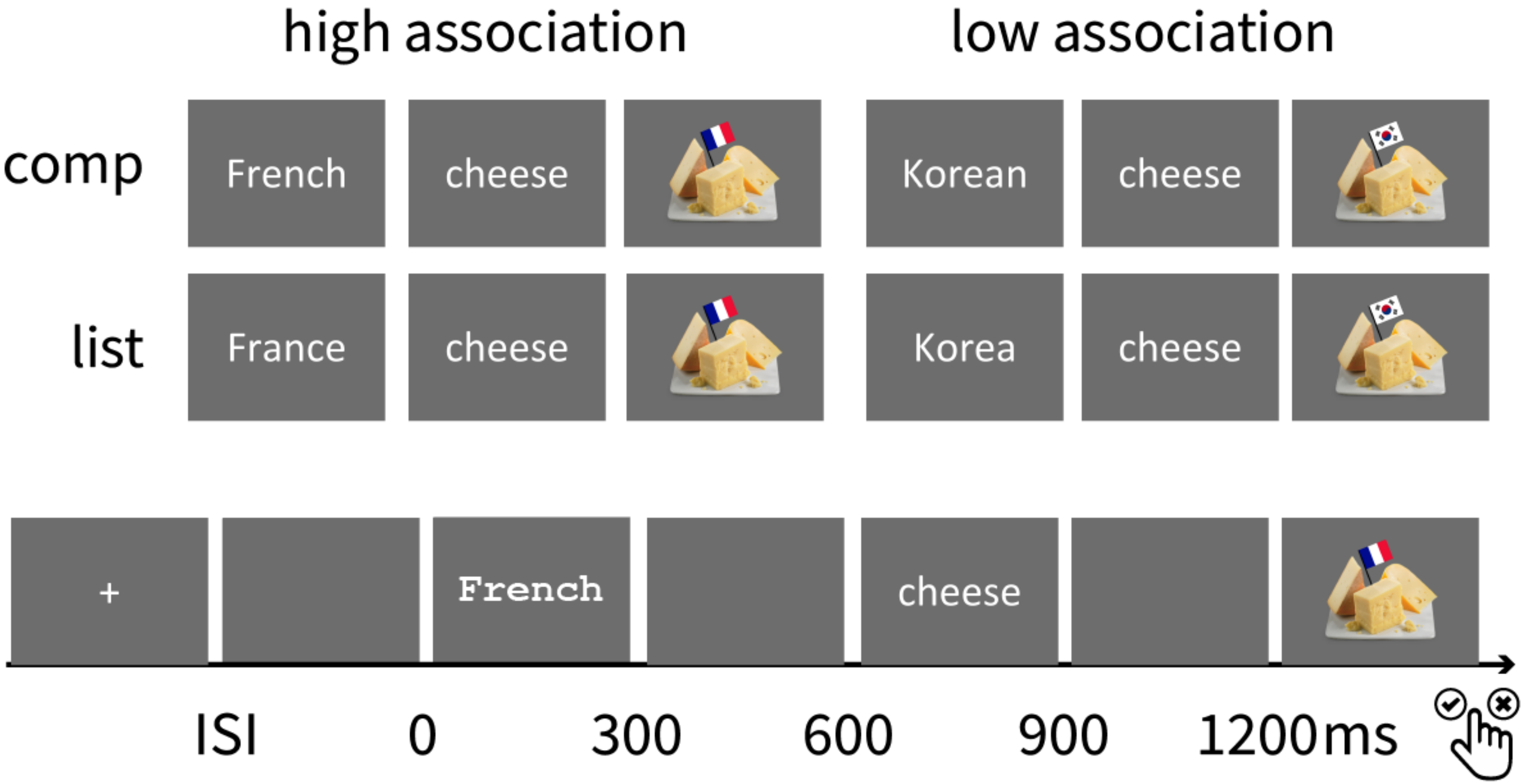
varying intercept only
y ~ x + (1|group)



varying slope and intercept
y ~ x + (x|group)

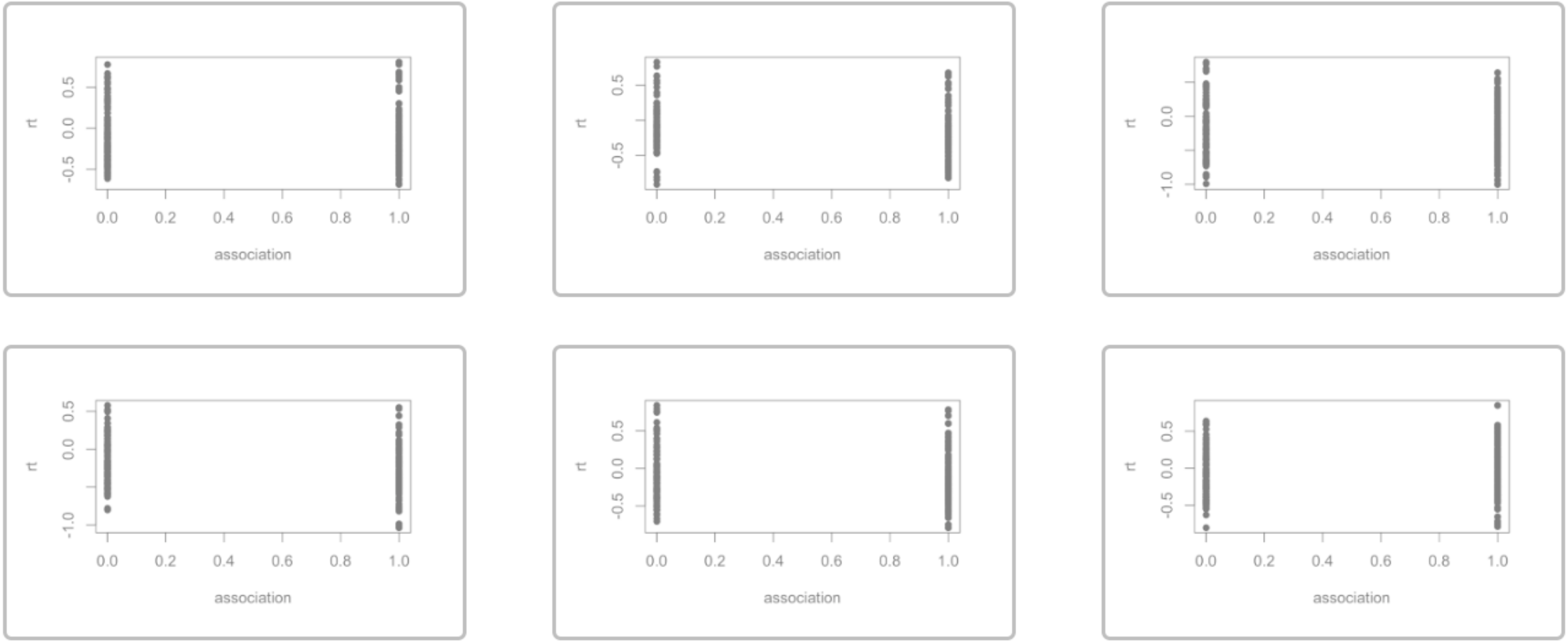


Example: Semantic composition vs. association



Each subject completed around 200 trials

Plot each subject's data



Multiple regression

```
md1 = lm(rt~composition+association+w1_freq+w2_freq,data=semassoc)
summary(md1)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.513117   0.089824   5.712 1.15e-08 ***
composition -0.004677   0.010589  -0.442   0.6587
association -0.026523   0.010573  -2.508   0.0121 *
w1_freq     -0.017687   0.004278  -4.134 3.60e-05 ***
w2_freq     -0.014376   0.002526  -5.692 1.30e-08 ***
```

LLM with random intercept

```r
install.packages('lmerTest')
library(lmerTest)
md2 = lmer(rt~composition+association+w1_freq+w2_freq+(1|subj),data=semassoc)
summary(md2)
```

```
Fixed effects:
              Estimate Std. Error        df t value Pr(>|t|)    
(Intercept)  5.091e-01  8.538e-02  1.041e+03   5.962 3.40e-09 ***
composition -6.339e-03  9.077e-03  7.725e+03  -0.698  0.48496    
association -2.693e-02  9.064e-03  7.725e+03  -2.971  0.00298 ** 
w1_freq     -1.716e-02  3.668e-03  7.725e+03  -4.678 2.95e-06 ***
w2_freq     -1.419e-02  2.165e-03  7.725e+03  -6.552 6.06e-11 ***
```

LLM with random intercept and slope

```
md3 = lmer(rt~composition+association+w1_freq+w2_freq+(association|subj),
data=semassoc)
summary(md3)
```

```
Fixed effects:
              Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  5.023e-01  8.395e-02  1.222e+03   5.983 2.88e-09 ***
composition -6.247e-03  9.015e-03  7.688e+03  -0.693   0.4884
association -2.935e-02  1.692e-02  4.256e+01  -1.734   0.0901 .
w1_freq     -1.682e-02  3.653e-03  7.718e+03  -4.606 4.18e-06 ***
w2_freq     -1.403e-02  2.151e-03  7.690e+03  -6.521 7.41e-11 ***
```

LLM with random intercept and slope

```
md4 = lmer(rt~composition+association+w1_freq+w2_freq+(association+composition|subj)
,data=semassoc,REML=F)
summary(md4)
```

```
boundary (singular) fit: see help('isSingular')

Fixed effects:
                Estimate Std. Error         df t value Pr(>|t|)    
(Intercept)   5.027e-01  8.330e-02  1.372e+03   6.035 2.05e-09 ***
composition  -5.548e-03  9.314e-03  2.893e+02  -0.596   0.5519    
association  -2.924e-02  1.674e-02  4.365e+01  -1.746   0.0878 .  
w1_freq      -1.688e-02  3.651e-03  7.714e+03  -4.622 3.86e-06 ***
w2_freq      -1.402e-02  2.150e-03  7.693e+03  -6.522 7.36e-11 ***
```

Model comparison

```
anova(md2,md3,md4)
```

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)	
md2	7	7968.2	8016.9	-3977.1	7954.2				
md3	9	7909.9	7972.5	-3946.0	7891.9	62.2936	2	2.973e-14	***
md4	12	7913.3	7996.8	-3944.7	7889.3	2.5879	3	0.4596	

Look for model with smaller AIC, BIC and larger log likelihood ratio