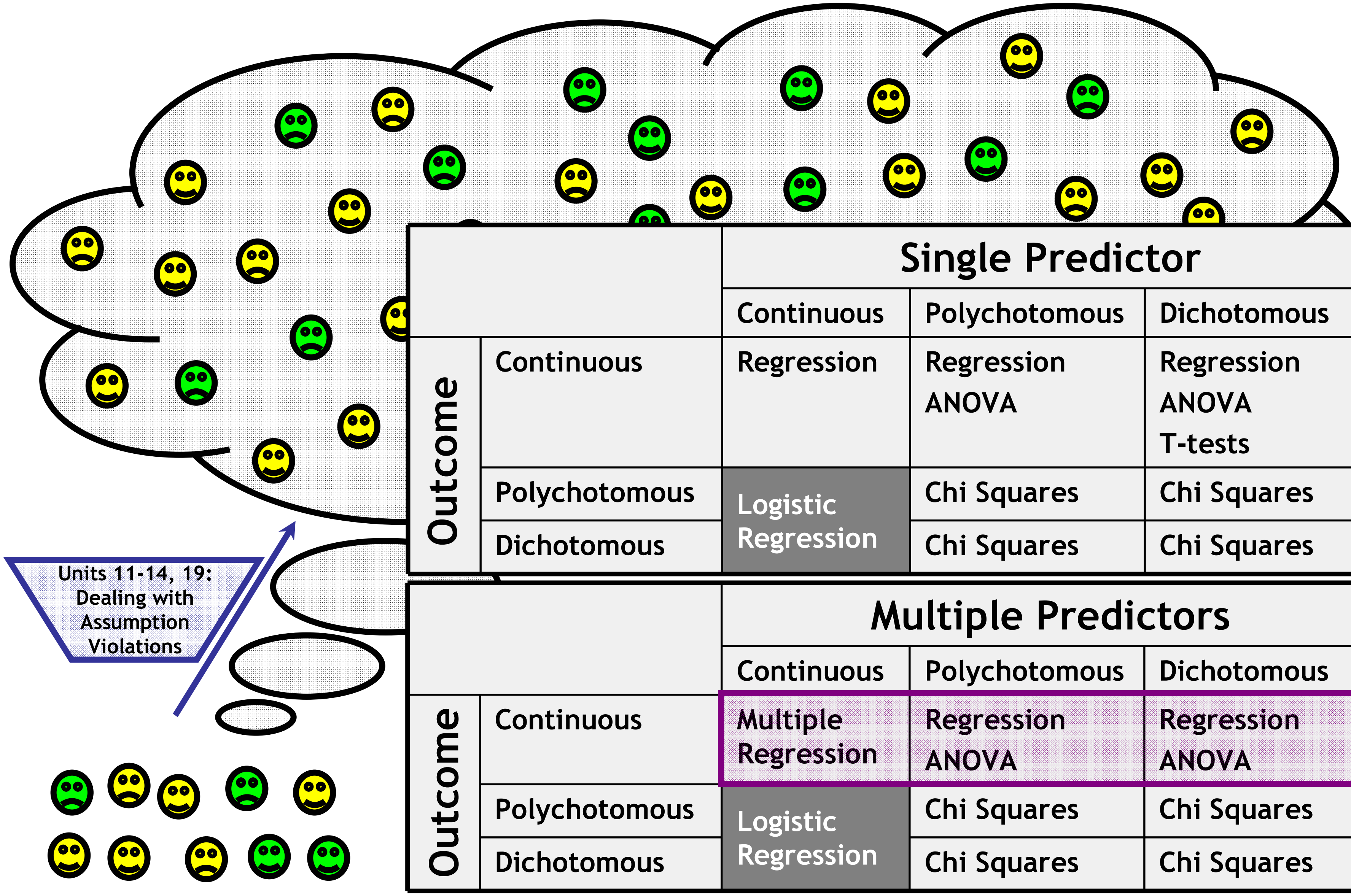
Unit 19: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):
READING, a continuous variable, test score, mean = 47 and standard deviation = 9
Predictor Variables (aka Independent Variables):
Question Predictor-
RACE, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White
Control Predictors-
HOMEWORK, hours per week, a continuous variable, mean = 6.0 and standard deviation = 4.7
FREELUNCH, a proxy for SES, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
ESL, English as a second language, a dichotomous variable, 1 = ESL, 0 = native speaker of English

- Unit 11: What is measurement error, and how does it affect our analyses?
- Unit 12: What tools can we use to detect assumption violations (e.g., ***outliers***)?
- Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?
- Unit 14: How do we deal with violations of the ***homoskedasticity*** assumption?
- Unit 15: What are the correlations among reading, race, ESL, and homework, controlling for SES?
- Unit 16: Is there a relationship between reading and race, controlling for SES, ESL and homework?
- Unit 17: Does the relationship between reading and race vary by levels of SES, ESL or homework?
- Unit 18: What are sensible strategies for building complex statistical models from scratch?
- Unit 19: How do we deal with violations of the ***independence*** assumption?

Unit 19: Road Map (Schematic)



Outcome	Single Predictor: Continuous	Single Predictor: Polychotomous	Single Predictor: Dichotomous
Continuous	Regression	Regression ANOVA	Regression ANOVA T-tests
Polychotomous	Logistic Regression	Chi Squares	Chi Squares
Dichotomous	Logistic Regression	Chi Squares	Chi Squares

Outcome	Multiple Predictors: Continuous	Multiple Predictors: Polychotomous	Multiple Predictors: Dichotomous
Continuous	Multiple Regression	Regression ANOVA	Regression ANOVA
Polychotomous	Logistic Regression	Chi Squares	Chi Squares
Dichotomous	Logistic Regression	Chi Squares	Chi Squares

In-Class Project

Repeated Measures Outcomes: *READING88 READING90 READING92*

Predictor: *RACE* (ASIAN, BLACK, LATINO, WHITE)

Fit and interpret a repeated measure ANOVA model:

- We will do this step-by-step together in class.

Fit and interpret this multilevel regression model:

$$READINGL_{ij} = \beta_0 + \beta_1 WAVE1_{ij} + \beta_1 WAVE2_{ij} + \beta_1 ASIAN_i + \beta_1 BLACK_i + \beta_1 LATINO_i$$
$$+ \beta_1 ASIANxWAVE1_{ij} + \beta_1 BLACKxWAVE1_{ij} + \beta_1 LATINOxWAVE1_{ij}$$
$$+ \beta_1 ASIANxWAVE2_{ij} + \beta_1 BLACKxWAVE2_{ij} + \beta_1 LATINOxWAVE2_{ij} + \varepsilon_{ij} + u_i$$

- We will restructure the data set together step-by-step in class.
- You will dummy code the variables by yourself but with as much help as you need.
- You will fit the model by yourself but with as much help as you need.
- We will interpret the results together step-by-step in class.

Unit 19: Funky Research Question

Theory: One group reads better than the other because…

Research Question: On average in the population, does Group 1 score higher on the reading test than Group 0?

Data Set: NELS (National Education Longitudinal Study) (n = 11856)

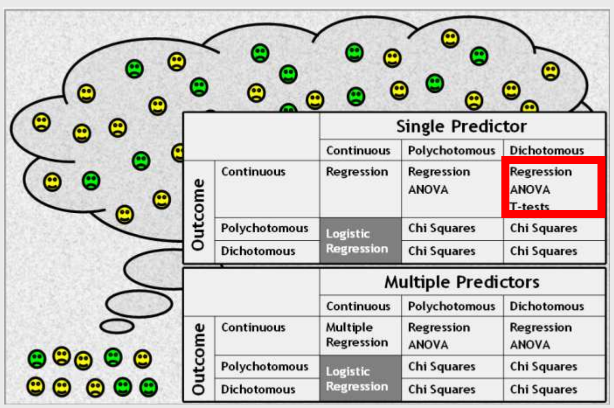
Variables:

Outcome: (***READINGL***) IRT Scaled Score on a Standardized Reading Test

Question Predictor: (***FUNKYVARIABLE***) A dichotomous variable indicating membership in one of two groups, Group 1 (***FUNKYVARIABLE*** = 1) or Group 0 (***FUNKYVARIABLE*** = 0)

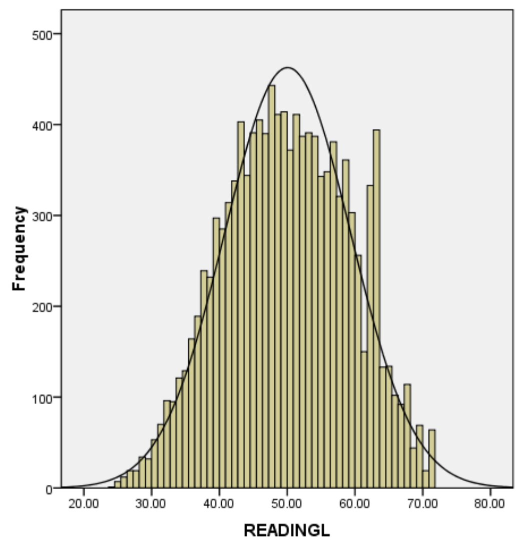
Model:

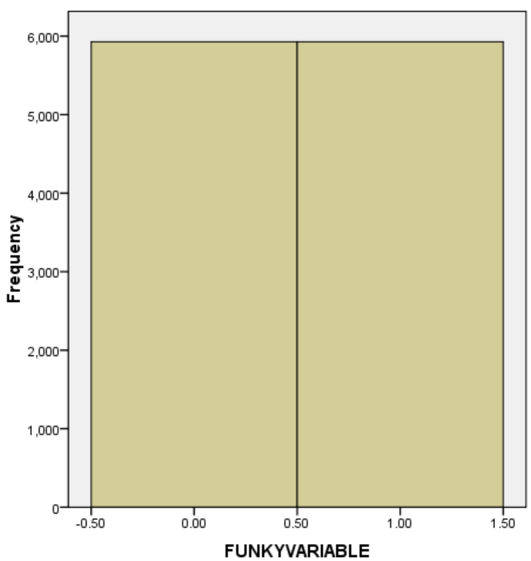
$$READINGL = \beta_0 + \beta_1 FUNKYVARIABLE + \varepsilon$$



We are going to answer this funkily abstract research question using the tools that we know and love. There is nothing new in this section. What makes this research question funky is my withholding of the meaning of *FUNKYVARIABLE*. If you get confused, you can replace in your mind *FUNKYVARIABLE* with *FEMALE*. So, instead of thinking about Group 1 and Group 0, you can think about females and males.

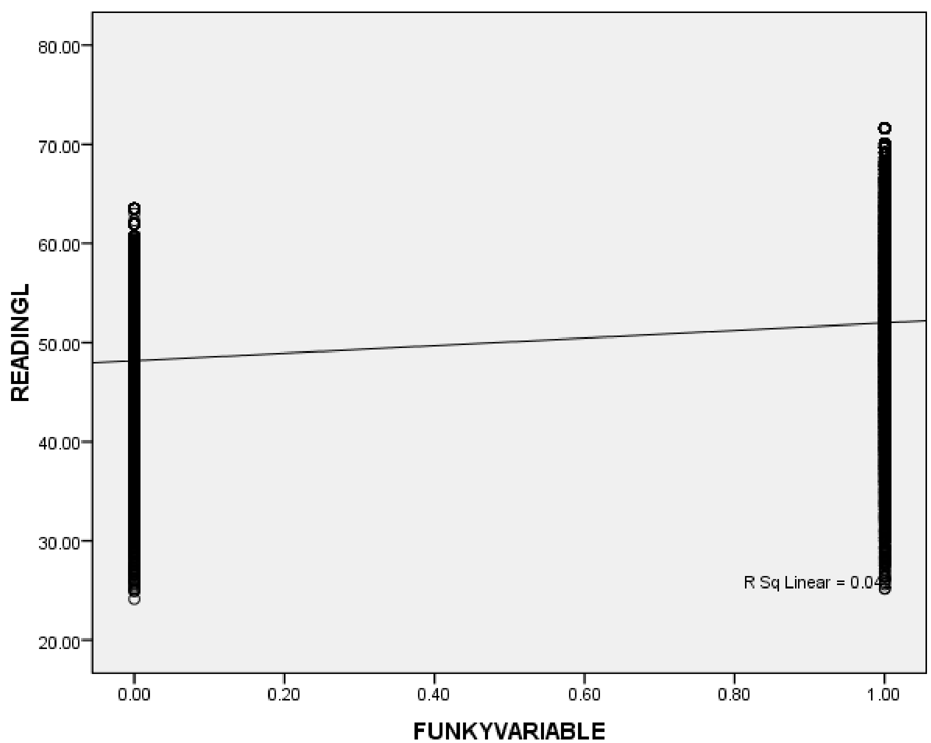
Exploratory Data Analysis





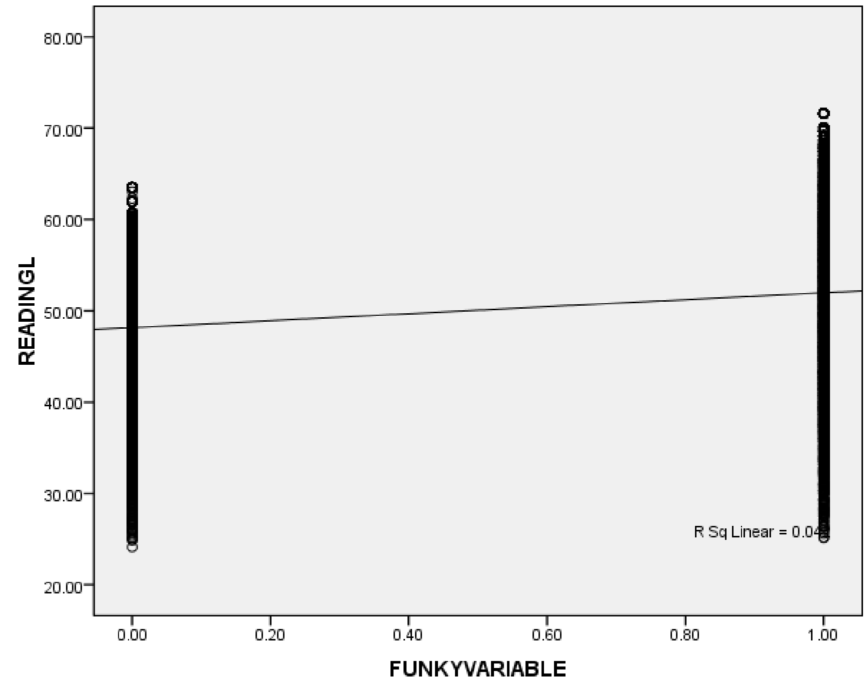
Statistics

		READINGL	FUNKYVARIABLE
N	Valid	11856	11856
	Missing	0	0
Mean		50.0695	.5000
Std. Deviation		9.29231	.50002
Minimum		24.14	.00
Maximum		71.61	1.00
Percentiles	25	43.2200	.0000
	50	49.9400	.5000
	75	57.1100	1.0000



Answering the Question Using Regression

On average in the population, Funky Group 1 tends to score higher than Funky Group 0 on the IRT scaled reading test, $t(11854) = 22.93$, $p < .001$. Based on 95% confidence intervals, we conclude that, in the population, the average score for Funky Group 1 ($M = 52.0$) is between 3.5 and 4.2 points higher than the average score for Funky Group 0 ($M = 48.2$)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.206[a]	.042	.042	9.09323

a. Predictors: (Constant), FUNKYVARIABLE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43473.909	1	43473.909	525.766	.000[a]
	Residual	980169.635	11854	82.687		
	Total	1023643.544	11855			

a. Predictors: (Constant), FUNKYVARIABLE

b. Dependent Variable: READINGL

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.155	.118		407.731	.000	47.923	48.386
	FUNKYVARIABLE	3.830	.167	.206	22.930	.000	3.502	4.157

a. Dependent Variable: READINGL

Answering the Question Using *t*-tests

t-test Output

Group Statistics

	FUNKY...	N	Mean	Std. Deviation	Std. Error Mean
READINGL	0	5928	48.1546	8.38109	.10885
	1	5928	51.9844	9.75351	.12668

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
READINGL	Equal variances assumed	170.159	.000	-22.930	11854	.000	-3.82979	.16702	-4.15719	-3.50240
	Equal variances not assumed			-22.930	11591.459	.000	-3.82979	.16702	-4.15719	-3.50240

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.206[a]	.042	.042	9.09323

a. Predictors: (Constant), FUNKYVARIABLE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43473.909	1	43473.909	525.766	.000[a]
	Residual	980169.635	11854	82.687		
	Total	1023643.544	11855			

a. Predictors: (Constant), FUNKYVARIABLE

b. Dependent Variable: READINGL

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.155	.118		407.731	.000	47.923	48.386
	FUNKYVARIABLE	3.830	.167	.206	22.930	.000	3.502	4.157

a. Dependent Variable: READINGL

Answering the Question Using ANOVA

ANOVA Output

Tests of Between-Subjects Effects

Dependent Variable:READINGL

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	43473.909[a]	1	43473.909	525.766	.000
Intercept	2.972E7	1	2.972E7	359457.530	.000
FUNKYVARIABLE	43473.909	1	43473.909	525.766	.000
Error	980169.635	11854	82.687		
Total	3.075E7	11856			
Corrected Total	1023643.544	11855			

a. R Squared = .042 (Adjusted R Squared = .042)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.206[a]	.042	.042	9.09323

a. Predictors: (Constant), FUNKYVARIABLE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43473.909	1	43473.909	525.766	.000[a]
	Residual	980169.635	11854	82.687		
	Total	1023643.544	11855			

a. Predictors: (Constant), FUNKYVARIABLE

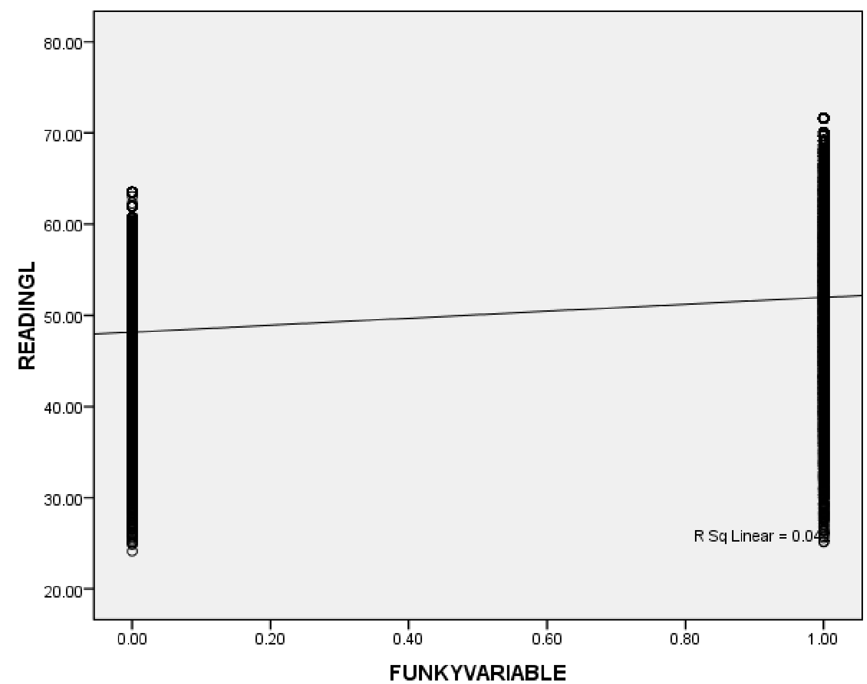
b. Dependent Variable: READINGL

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.155	.118		407.731	.000	47.923	48.386
	FUNKYVARIABLE	3.830	.167	.206	22.930	.000	3.502	4.157

a. Dependent Variable: READINGL

Important Observations/Reminders



The intercept (aka, constant) is going to play a very important role in things to come. Recall that the intercept is the mean of our reference category.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.206[a]	.042	.042	9.09323

a. Predictors: (Constant), FUNKYVARIABLE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43473.909	1	43473.909	525.766	.000[a]
	Residual	980169.635	11854	82.687		
	Total	1023643.544	11855			

a. Predictors: (Constant), FUNKYVARIABLE

b. Dependent Variable: READINGL

Note that the F statistic is simply the square of the t statistic (in simple linear regression).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.155	.118		407.731	.000	47.923	48.386
	FUNKYVARIABLE	3.830	.167	.206	22.930	.000	3.502	4.157

a. Dependent Variable: READINGL

Thinking More Deeply About the y-Intercept

Model 1:

$$READINGL = \beta_0 + \beta_1 FUNKYVARIABLE + \varepsilon$$

The y-intercept is represented by β_0, which in turn represents the mean of *READINGL* when all the predictors have values of zero. What does β_0 represent when there are no predictors in the model?

Statistics

READINGL

N	Valid	11856
	Missing	0
	Mean	50
	Variance	86

Model 0:

$$READINGL = \beta_0 + \varepsilon$$

When there are no predictors in the model, β_0 represents the (unconditional) mean of *READINGL*. Recall that in the absence of further information, the mean is our best guess for individuals, but we recognize that the guess is in all probability wrong by a certain amount, so we make sure that we have an error term in our model, ε.

Variance (i.e., the average squared mean deviation) is a measure of how wrong the mean is as a predictor of individuals.

Output from Fitting the Unconditional Model (Model 0)

Note that SPSS does not allow us to fit unconditional OLS regression models, so I made this output by hand.

Mean *READINGL*

Standard Deviation

Variance

Sum of Squared Mean Deviations

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
0	.000[a]	.000	.000	9.29231

a. Predictors: (Constant)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
0	Regression	0	0	0	0	1.00[a]
	Residual	1023643.544	11855	86.347		
	Total	1023643.544	11855			

a. Predictors: (Constant)

b. Dependent Variable: READINGL

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
0	(Constant)	50.070	.08534		586.703	.000	49.902	50.237

a. Dependent Variable: READINGL

Output from Fitting the Conditional Model (Model 1)

In this model, we make predictions of our outcome conditional on our predictor, which is *status quo* for us.

Mean *READINGL* Conditional on *FUNKYVARIABLE*=0

Standard Deviation* of the Residuals

Variance* of the Residuals

Sum of Squared Residuals (i.e., Deviations From the Regression Line)

* Basically

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.206[a]	.042	.042	9.09323

a. Predictors: (Constant), FUNKYVARIABLE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43473.909	1	43473.909	525.766	.000[a]
	Residual	980169.635	11854	82.687		
	Total	1023643.544	11855			

a. Predictors: (Constant), FUNKYVARIABLE

b. Dependent Variable: READINGL

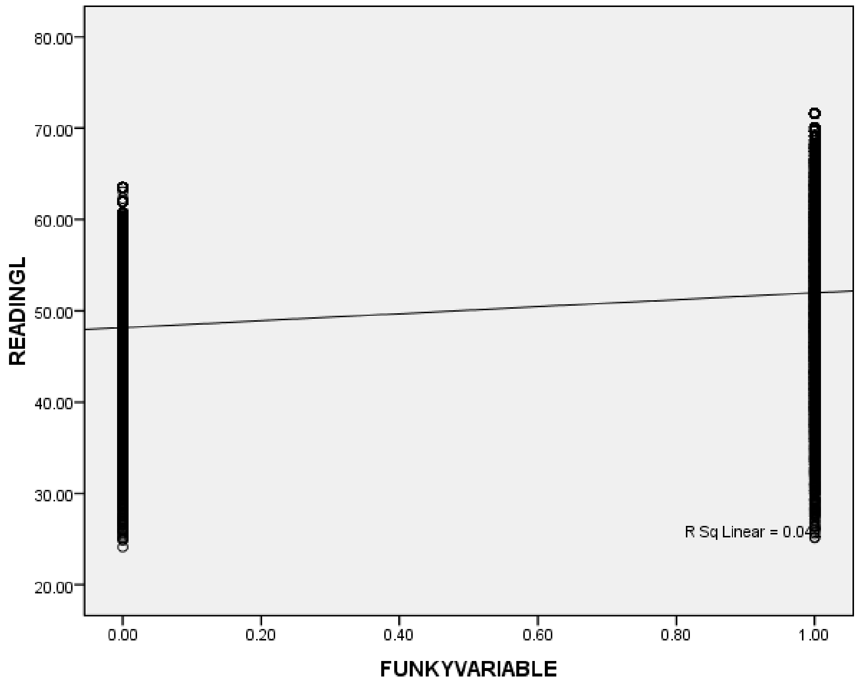
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.155	.118		407.731	.000	47.923	48.386
	FUNKYVARIABLE	3.830	.167	.206	22.930	.000	3.502	4.157

a. Dependent Variable: READINGL

Checking Assumptions for Model 1: Searching HI-N-LO

- **Heteroskedasticity**—We can judge by looking at the right graphs.
- **Independence**—We cannot judge by looking at any graphs. We need to understand our sample and our variable(s).
- **Normality**—We can judge by looking at the right graphs.
- **Linearity**—We can judge by looking at the right graphs.
- **Outliers**—We can judge by looking at the right graphs.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.206[a]	.042	.042	9.09323

a. Predictors: (Constant), FUNKYVARIABLE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43473.909	1	43473.909	525.766	.000[a]
	Residual	980169.635	11854	82.687		
	Total	1023643.544	11855			

a. Predictors: (Constant), FUNKYVARIABLE

b. Dependent Variable: READINGL

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.155	.118		407.731	.000	47.923	48.386
	FUNKYVARIABLE	3.830	.167	.206	22.930	.000	3.502	4.157

a. Dependent Variable: READINGL

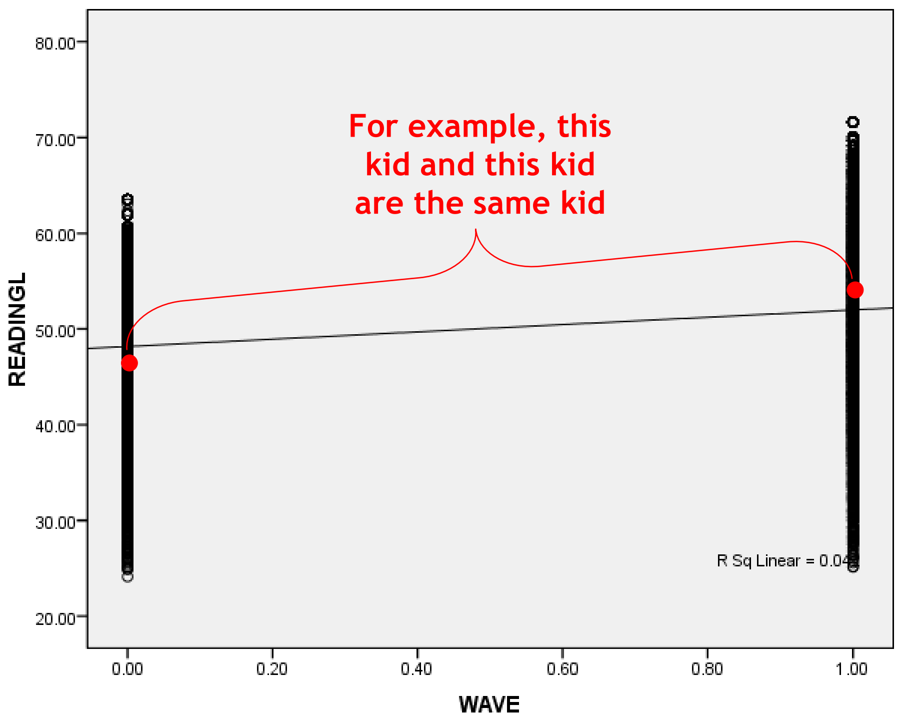
But, what is our variable?

Riddle Revealed

Funky Question Predictor: (***FUNKYVARIABLE***) A dichotomous variable indicating membership in one of two groups, Group 1 (***FUNKYVARIABLE*** = 1) or Group 0 (***FUNKYVARIABLE*** = 0)

Real Question Predictor: (***WAVE***) A dichotomous variable indicating the wave in which the reading test was taken, the baseline test was taken in 1988, the 8th grade, (***WAVE*** = 0) and the follow-up test was taken in 1990, the 10th grade (***WAVE*** = 1).

We have 11856 observations but only 5928 subjects (with two observations per subject, a baseline observation and a follow-up observation).



A new (multilevel) way of thinking:

Scores Nested in Students

Students Nested in Classrooms

Classrooms Nested In Schools

Schools Nested in Districts

Districts Nested in States

Children Nested in Families

Families Nested in Neighborhoods

Neighborhoods Nested in Cities

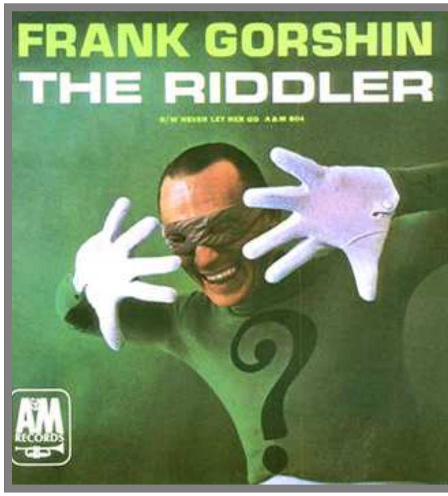
Babies Nested in Nurseries

Nurseries Nested in Hospitals

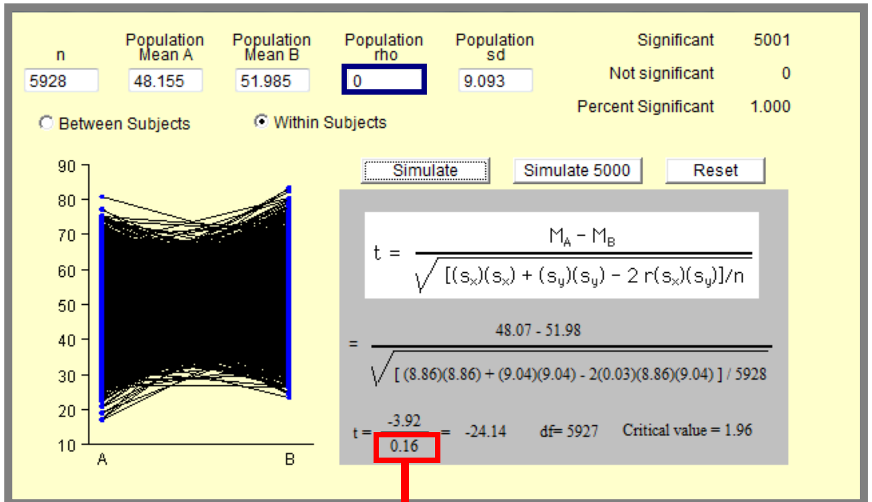
We will learn to handle two levels at a time:

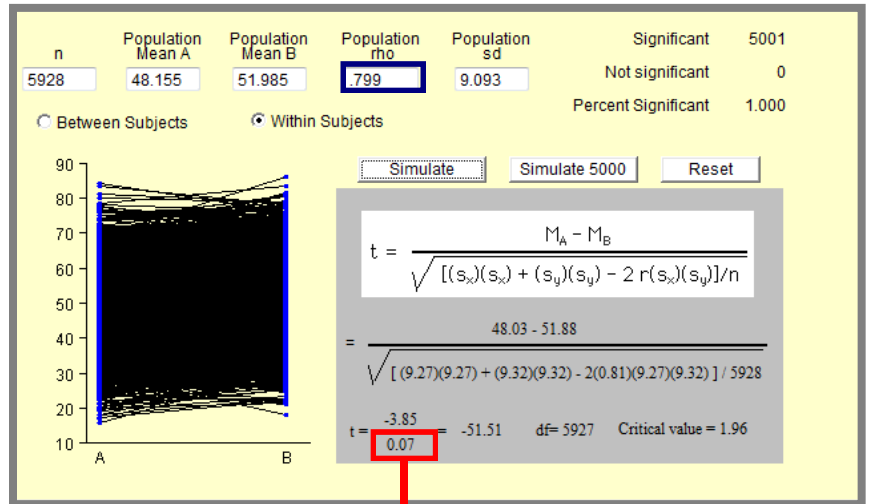
“Observations” Nested in “Clusters”

Our observations are clustered (in pairs); thus, our independence assumption is violated.



Independence Schmindependence: Why Care?





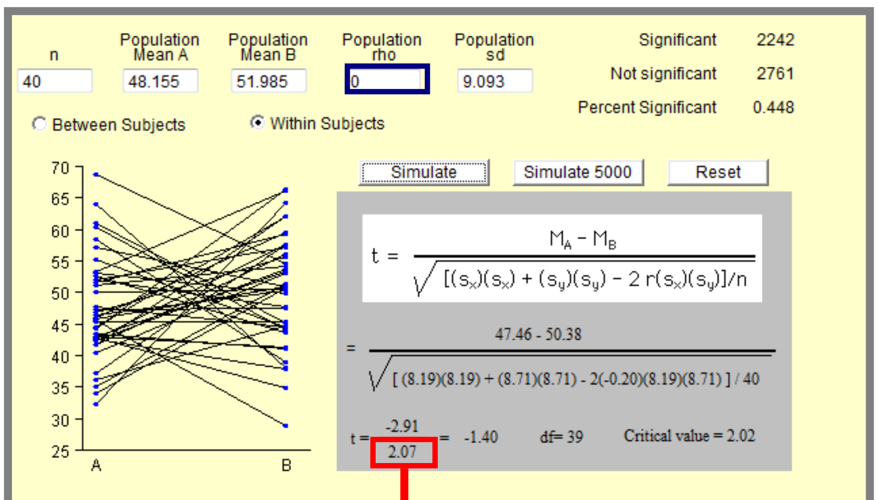
Mistaking order for chaos is no way to go about the business of truth.

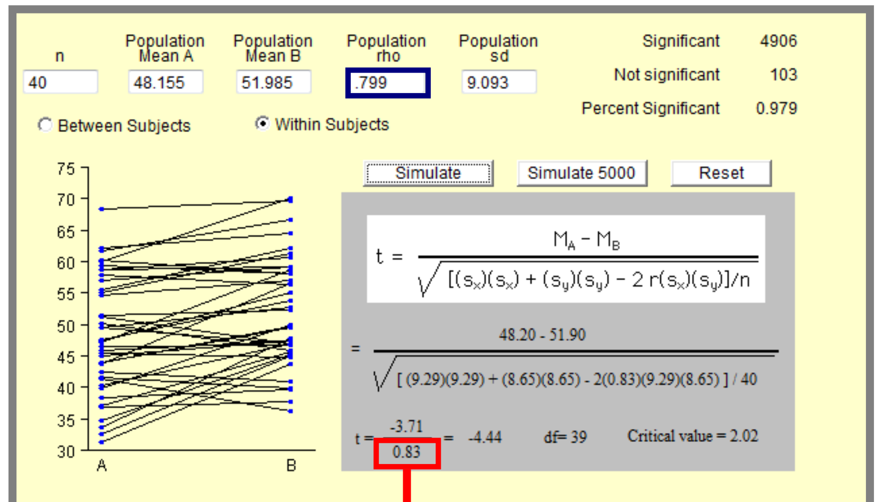
Note the Correlations

Note the Standard Errors

http://onlinestatbook.com/stat_sim/repeated_measures/index.html

Even when a huge sample size makes statistical significance a foregone conclusion, we still want the right standard errors for our confidence intervals.





One Final Riddle Before We Get Started

Riddle: A class of students takes a midterm exam and a final exam. The average score on the midterm exam is 78, and the average score on the final exam is 92. What is the correlation between the two sets of exam scores? Can you say exactly? Can you at least say the direction?

Answer: We have no clue! If you are like me, your intuition is that the correlation *must* be positive, but it *could* be negative. Imagine if all the people who did the worst on the midterm exam were jarred into working harder (and smarter), so they ended up doing the best on the final exam.

	Name	Midterm	Final
1	Sorastro	90.00	89.00
2	Tamino	86.00	90.00
3	Papageno	82.00	91.00
4	Astrofiammante	78.00	92.00
5	Pamina	74.00	93.00
6	Monostatos	70.00	94.00
7	Papagena	66.00	95.00

In this data set (n = 7), there is a perfect negative correlation between the midterm scores and the final scores. The means are different (*M* = 78 and *M* = 92), and the standard deviations also happen to be different (*SD* = 8.6 and *SD* = 2.2). But, the correlation does not care! I teach the correlation coefficient as the slope coefficient from the regression of a standardized outcome on a standardized predictor. When we standardize, we force the means to be zero and the standard deviations to be one so that we can compare apples to apples. See Unit 4 for a refresher. Algebraically, a correlation is the average of the products of the z-scores:

$$r_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

We subtract out the means and divide away the standard deviations.

Unit 19: Most Basic Research Question

Theory: Students improve their reading skills from the 8th grade to the 10th grade.

Research Question: On average in the population, do students improve on the reading test from the 8th grade to the 10th grade? If so, by how much do they improve?

Because this research question is so basic, we have a wide choice of tools: paired samples t-tests, repeated measures ANOVA, and multilevel regression modeling. We will try all three in order from simple (and least flexible) to complicated (and most flexible).

Data Set: NELS (National Education Longitudinal Study) (n = 5928)

Variables:

Outcome: (***READINGL***) IRT Scaled Scores on a Standardized Reading Test

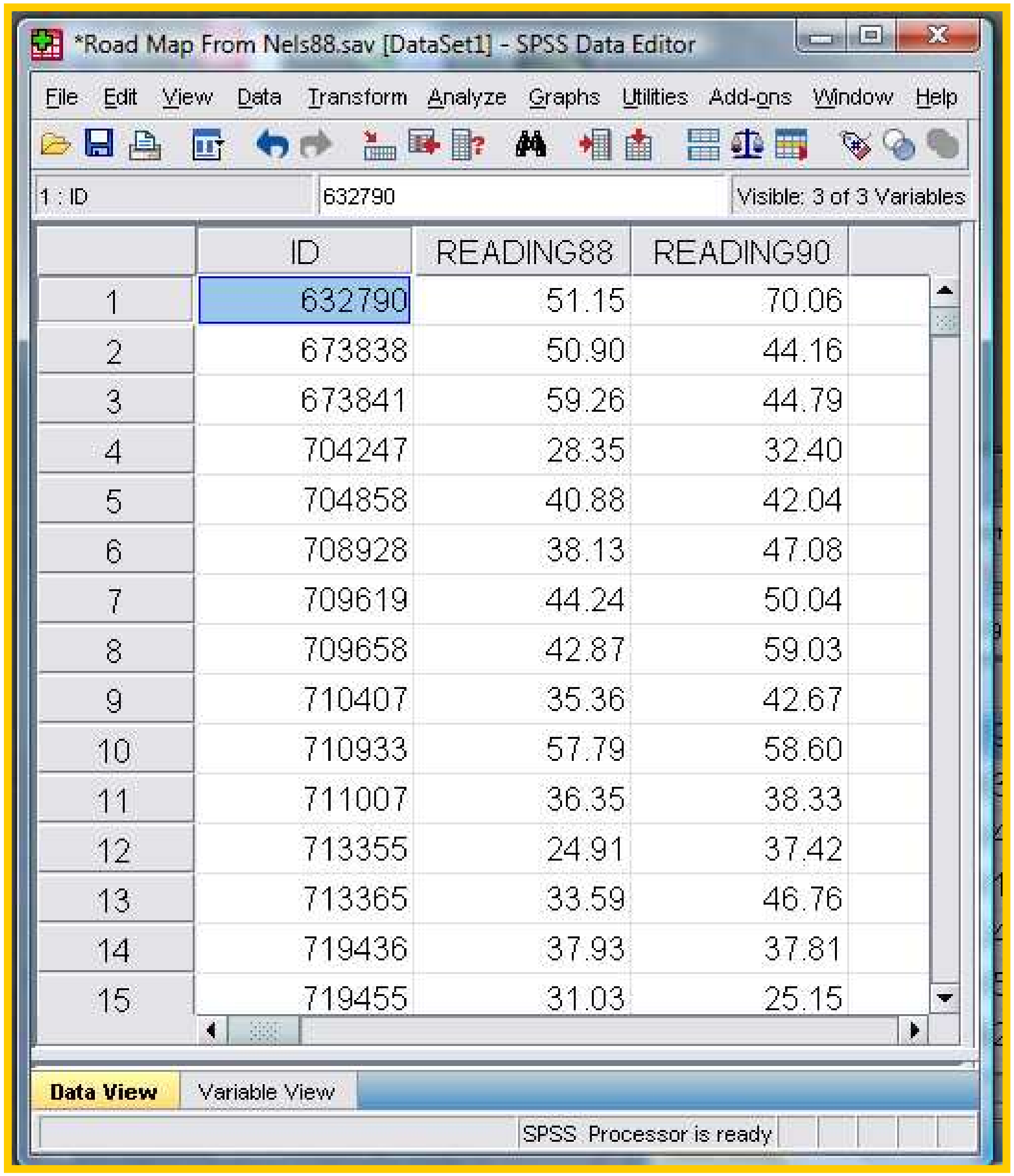
Question Predictor: **From the t-test perspective there is no real predictor, just two (paired) samples. From the ANOVA perspective there is no real predictor, just a single repeated measures factor, a sort of fusion of our outcome information and wave information. However, from the regression perspective, we get to think in terms of outcomes and predictors and apply all our model building strategies:**

(***WAVE***) A dichotomous variable indicating the wave in which the test was taken where ***WAVE*** = 0 denotes the baseline, 8th grade, 1988 scores and ***WAVE*** = 1 denotes the follow-up, 10th grade, 1990 scores.

Regression Model: $$READINGL_{ij} = \beta_0 + \beta_1 WAVE_{ij} + \varepsilon_{ij} + u_i$$

Notice the *ij* subscripts and a second type of error

Data Set (For Paired Samples *t*-test and Repeated Measures ANOVA)



	ID	READING88	READING90
1	632790	51.15	70.06
2	673838	50.90	44.16
3	673841	59.26	44.79
4	704247	28.35	32.40
5	704858	40.88	42.04
6	708928	38.13	47.08
7	709619	44.24	50.04
8	709658	42.87	59.03
9	710407	35.36	42.67
10	710933	57.79	58.60
11	711007	36.35	38.33
12	713355	24.91	37.42
13	713365	33.59	46.76
14	719436	37.93	37.81
15	719455	31.03	25.15

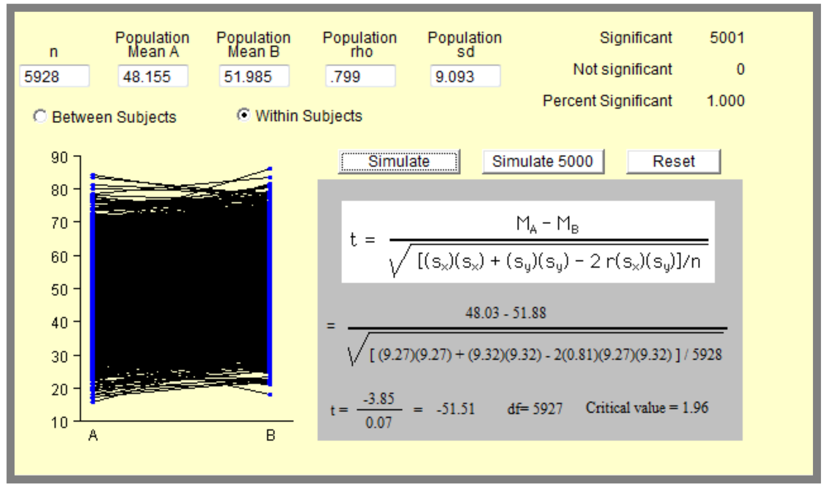
This data structure is very familiar to us. Rows represent kids. We see that the first kid in our data set has 632790 for an ID number and scores 51.15 points on the 1988 (8th grade, baseline) reading test and 70.06 points on the 1990 (10th grade, follow-up) reading test. Columns represent variables. We have an ID variable to help us identify kids, and we have two test-score variables.

For multilevel regression modeling, we will need to restructure this data set into a “person-period data set.” But, no worries, because SPSS will basically do the work for us. For now, however, while we work through *t*-tests and ANOVAs, we’ll stay in this familiar territory.

t-test Perspective

Standard errors come in many flavors, but at their core they are just special standard deviations; they are standard deviations of sampling distributions. The bigger the sample size, the smaller the standard deviation of the sampling distribution, so we estimate standard errors by dividing our observed standard deviations by the square root of our sample sizes. See Unit 6 for a refresher. There are slight twists for different tests, and the twist here is that we take into consideration the correlation.

$$t = \frac{M_A - M_B}{\sqrt{[(s_x)(s_x) + (s_y)(s_y) - 2r(s_x)(s_y)]/n}}$$



http://onlinestatbook.com/stat_sim/repeated_measures/index.html

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	READING IRT THETA 1988	48.1546	5928	8.38109	.10885
	READING IRT THETA 1990	51.9844	5928	9.75351	.12668

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	READING IRT THETA 1988 & READING IRT THETA 1990	5928	.799	.000

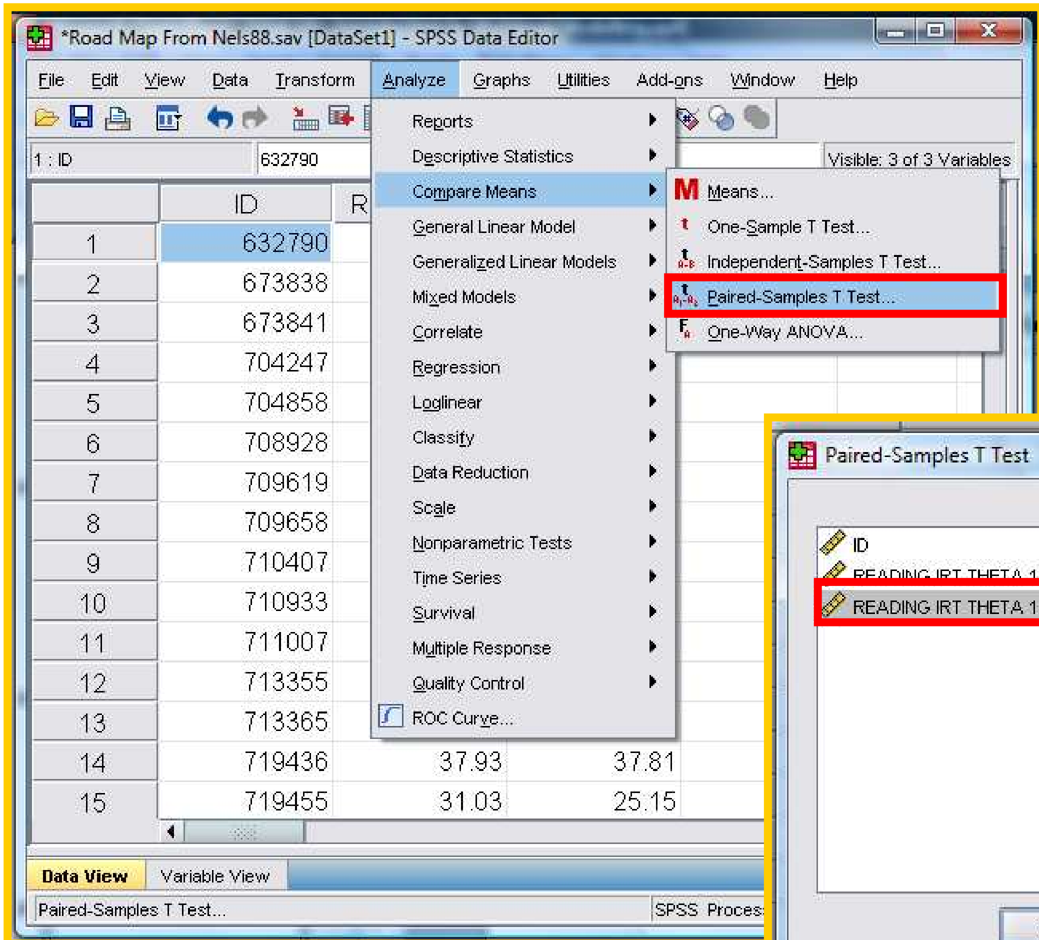
Take some time to work through this. Here is a spot where the algebra can be insightful. For example, we know that a large sample size is good. See how the samply goodness of the size works into the equation.

Not that when the correlation is zero, the entire $-2r(s_x)(s_y)$ is zeroed out, and we end up with a run-of-the-mill *t*-test.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference: Lower	95% Confidence Interval of the Difference: Upper			
Pair 1	READING IRT THETA 1988 - READING IRT THETA 1990	-3.82979	5.88792	.07647	-3.97971	-3.67988	-50.080	5927	.000

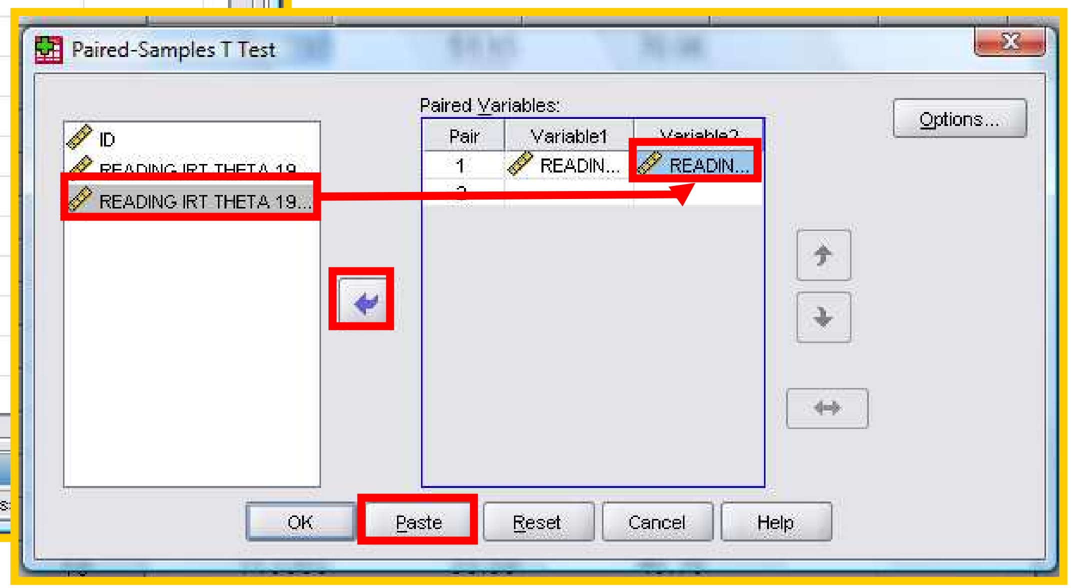
Paired Samples *t*-tests in SPSS



Go to Analyze > Compare Means > Paired-Samples T Test…

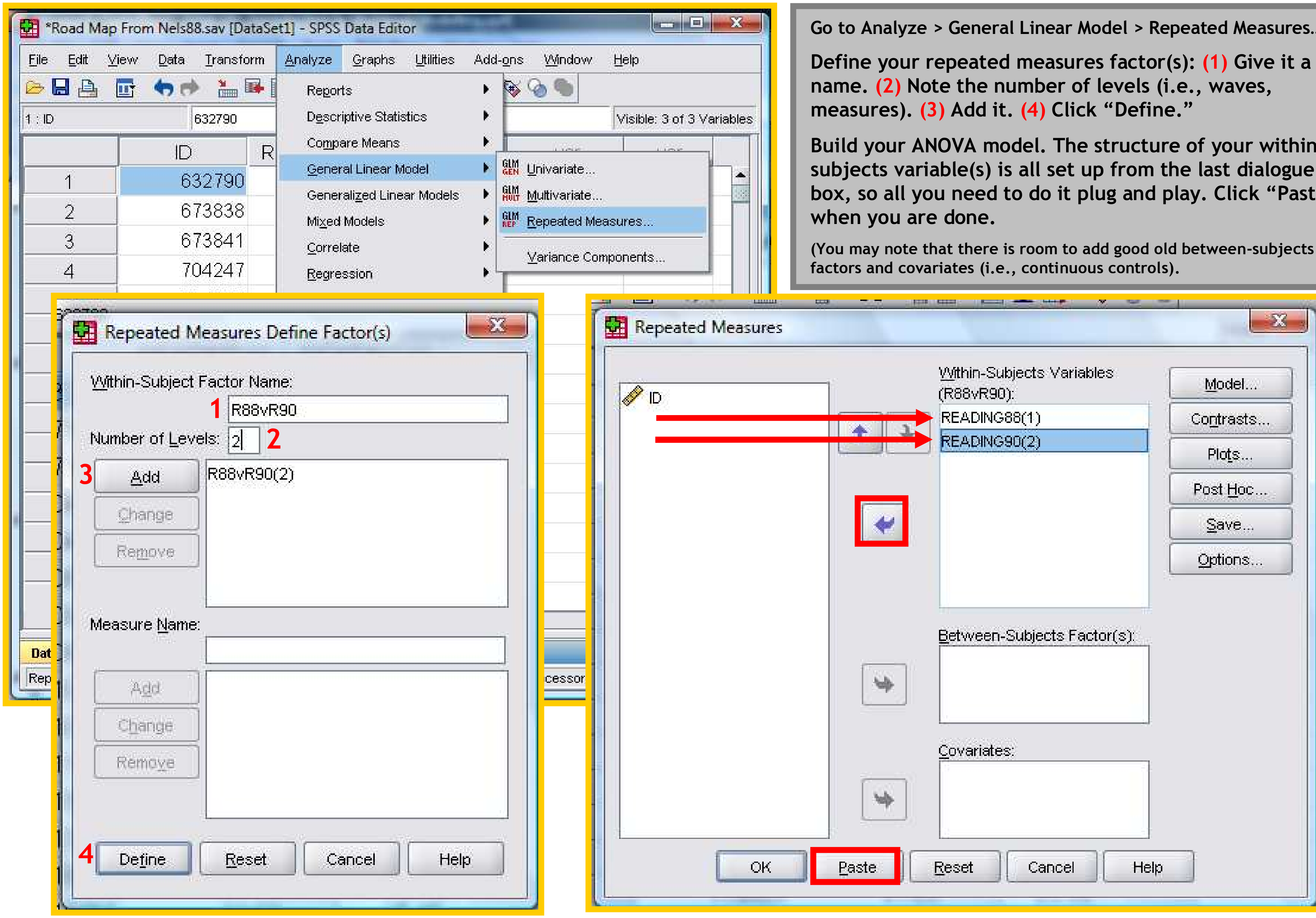
Select your first measure and assign it to the Variable 1 column, and select your second measure and assign it to the Variable 2 column (shown).

Click past when you are done, and run your syntax.




```
T-TEST PAIRS=READING88 WITH READING90 (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

Repeated Measures ANOVA in SPSS



Go to Analyze > General Linear Model > Repeated Measures…

Define your repeated measures factor(s): (1) Give it a name. (2) Note the number of levels (i.e., waves, measures). (3) Add it. (4) Click "Define."

Build your ANOVA model. The structure of your within-subjects variable(s) is all set up from the last dialogue box, so all you need to do it plug and play. Click "Paste" when you are done.

(You may note that there is room to add good old between-subjects factors and covariates (i.e., continuous controls).

ANOVA Perspective

The syntax is fairly simple, and the output should be very simple, but SPSS produces a crap load of distracting output. Much of the distracting output has to do with the sphericity assumption, which you can read about in Chapter 13 of the OnlineStatBook.Com. Of the umpteen tables, this is the only really important table, and still it's cluttered with junk. It should only be two lines:

```
GLM READING88 READING90
  /WSFACTOR=R88vsR90 2 Simple
  /METHOD=SSTYPE(3)
  /CRITERIA=ALPHA(.05)
  /WSDESIGN=R88vsR90.
```

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
R88vsR90	Sphericity Assumed	43473.909	1	43473.909	2508.042	.000
	Greenhouse-Geisser	43473.909	1.000	43473.909	2508.042	.000
	Huynh-Feldt	43473.909	1.000	43473.909	2508.042	.000
	Lower-bound	43473.909	1.000	43473.909	2508.042	.000
Error(R88vsR90)	Sphericity Assumed	102737.452	5927	17.334		
	Greenhouse-Geisser	102737.452	5927.000	17.334		
	Huynh-Feldt	102737.452	5927.000	17.334		
	Lower-bound	102737.452	5927.000	17.334		

Recall that the *F* statistic is the square of the *t* statistic. The t statistic from our paired samples *t*-test was -50.08.

$-50.08^2 = 2508.043$

In ANOVA, the correlation gets worked in through the mean squares. (And, that's all we really need to know.)

We conducted a one-way within-subjects ANOVA to determine whether IRT scales reading scores improved from 8th grade to 10th grade in the population of U.S. school children of the late '80s and early '90s. We observe a statistically significant *F* value, $F(1, 5927) = 2508.04$, $p < .001$, partial $\eta^2 = .28$. A comparison of means suggests that students on average improved 3.83 points from the 1988 8th grade reading test ($M = 48.15$, $SD = 8.38$) to the 1990 10th grade reading test ($M = 51.98$, $SD = 9.75$).

As always with ANOVA, we need to use planned contrasts, graphical plots, *post hoc* tests, and other options to get the juicy details.

Regression Perspective

Model 1:

$$READINGL_{ij} = \beta_0 + \beta_1 WAVE_{ij} + \varepsilon_{ij} + u_i$$

This looks very much like the regression models with which we have been working all along the way. The only differences are that now we have *ij* subscripts and a second error term. In the next few slides, we will examine the two differences and their implications.

Note that the subscript issue is really just a picky detail, but I want to emphasize it in order to get us thinking about cluster-observation data structure. In particular, we want to think about student-score data structures (aka, person-period data structures) for our research question. For other research questions, we may want to think about mother-child data structures or school-student data structures.

The magic of multilevel regression modeling happens in the complex error term: we have one error term for the observation level and another error term for the cluster level. In our example, we will have student-level error and score-level error. The key to parsing the error will be the unconditional model:

Model 0:

$$READINGL_{ij} = \beta_0 + \varepsilon_{ij} + u_i$$

Or, equivalently:

$$READINGL_{ij} = (\beta_0 + u_i) + \varepsilon_{ij}$$

Regression Perspective: *ij* Subscripts

Model 1:

$$READINGL_{ij} = \beta_0 + \beta_1 WAVE_{ij} + \varepsilon_{ij} + u_i$$

We use *ij* subscripts to distinguish our observation-level variables from our cluster-level variables. Observation-level variables get an *ij* subscript. Cluster-level variables get simply an *i* subscript.

In the problem at hand, we have scores (i.e., our observations) nested within students (i.e., our clusters). However, the system we are going to develop is flexible enough to handle any two-level nested structure. For example, we might have children (i.e., our observations) nested within mothers (i.e., our clusters), or we might have students (i.e., our observations) nested within schools (i.e., our clusters).

$WAVE_{ij}$ represents the value of the *WAVE* variable for the j^{th} score of the i^{th} student. E.g., for the 2^{nd} score of the 896^{th} student, *WAVE* = 1.

$READINGL_{ij}$ represents the value of the *READINGL* variable for the j^{th} score of the i^{th} student. E.g., for the 2^{nd} score of the 896^{th} student, *READINGL* = 61.

Model 2:

$$READINGL_{ij} = \beta_0 + \beta_1 WAVE_{ij} + \beta_2 ASIAN_i + \beta_3 BLACK_i + \beta_4 LATINO_i + \varepsilon_{ij} + u_i$$

$ASIAN_i$ represents the value of the *ASIAN* variable for the i^{th} student. E.g., for the 896^{th} student, *ASIAN* = 0. (Note that since this is a student-level variable, there is no need to attach it to a particular score.)

Regression Perspective: *ij* Subscripts (More Examples)

This is a study in which we ask whether smarter mother's have heavier newborns, controlling for length of gestation. We do not want to ignore the fact that newborns are nested within mothers, because we have twins and other sibs in our study.

Model X:

$$BIRTHWEIGHT_{ij} = \beta_0 + \beta_1 MOMIQ_i + \beta_2 GESTATION_{ij} + \varepsilon_{ij} + u_i$$

$MOMIQ_i$ represents the value of the *MOMIQ* variable for the i^{th} mother. E.g., for the 57^{th} mother, *MOMIQ* = 105.

$GESTATION_{ij}$ represents the value of the *GESTATION* variable for the j^{th} child of the i^{th} mother. E.g., for the 3^{rd} child of the 57^{th} mother, *GESTATION* = 271.

This is a study in which we ask about the Black/White math achievement gap and whether it varies by the racial composition of schools.

Model Y:

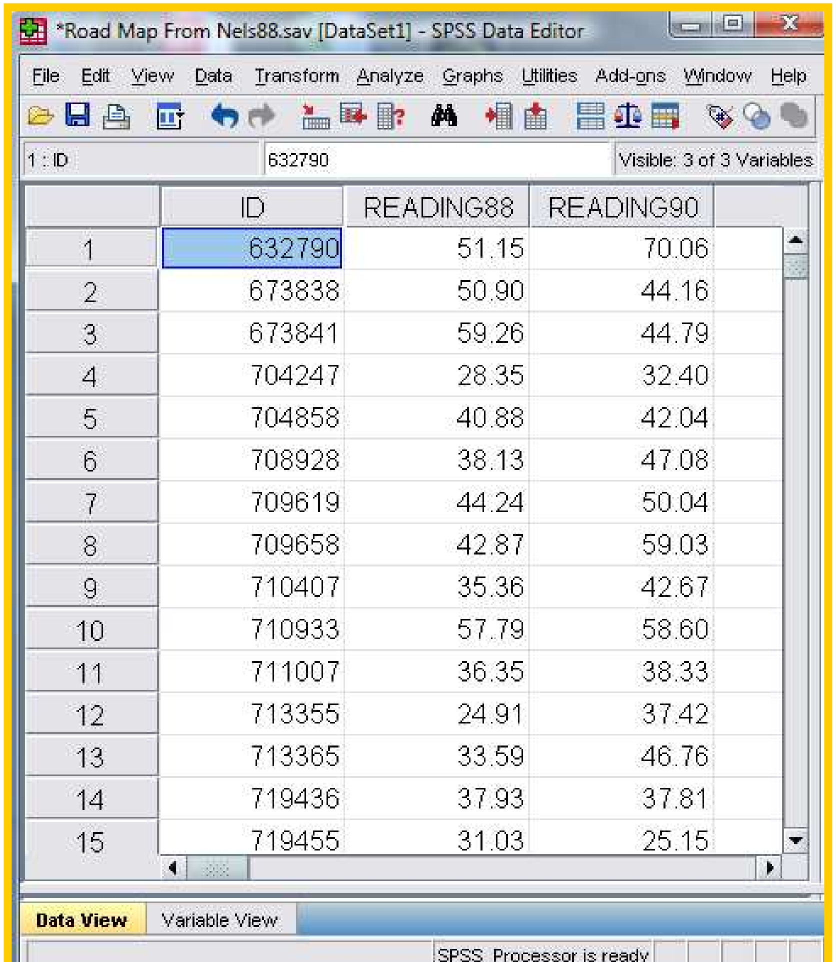
$$MATH_{ij} = \beta_0 + \beta_1 BLACK_{ij} + \beta_2 BWRATIO_i + \beta_1 BLACKxBWRATIO_{ij} + \varepsilon_{ij} + u_i$$

$BLACK_{ij}$ represents the value of the *BLACK* variable for the j^{th} student of the i^{th} school. E.g., for the 83^{rd} student of the 5^{th} school, *BLACK* = 1.

$BWRATIO_i$ represents the value of the *BWRATIO* variable for the i^{th} SCHOOL. E.g., for the 5^{th} school, *BWRATIO* = 0.75.

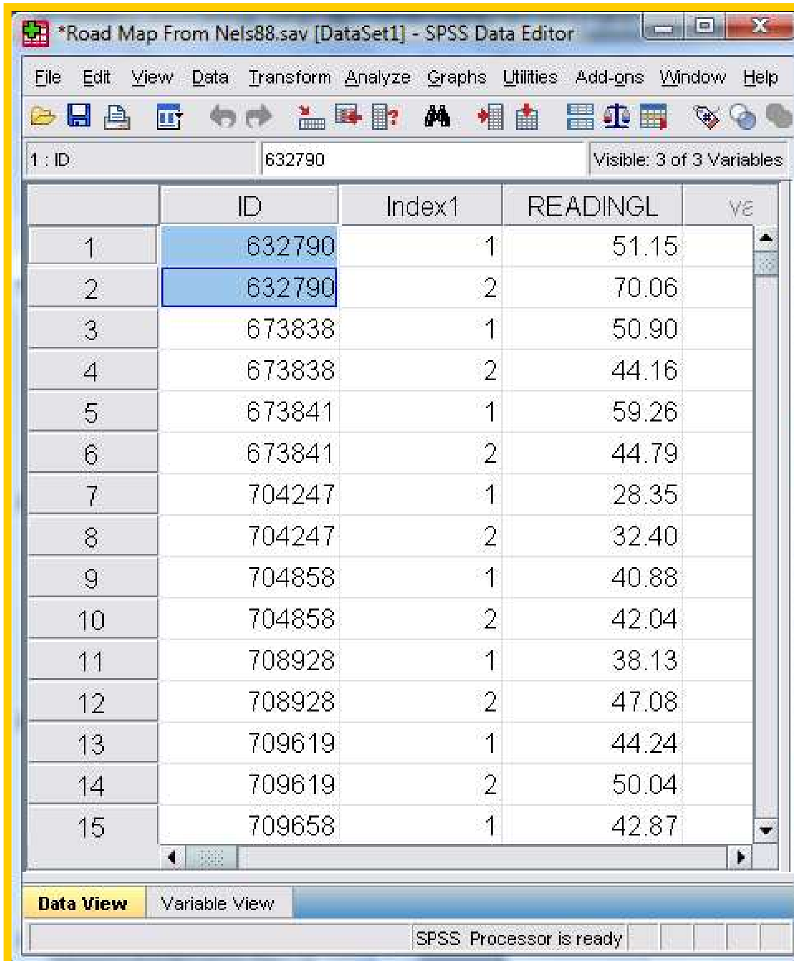
Person-Period Data Set Structure

Old Structure



	ID	READING88	READING90
1	632790	51.15	70.06
2	673838	50.90	44.16
3	673841	59.26	44.79
4	704247	28.35	32.40
5	704858	40.88	42.04
6	708928	38.13	47.08
7	709619	44.24	50.04
8	709658	42.87	59.03
9	710407	35.36	42.67
10	710933	57.79	58.60
11	711007	36.35	38.33
12	713355	24.91	37.42
13	713365	33.59	46.76
14	719436	37.93	37.81
15	719455	31.03	25.15

New Structure

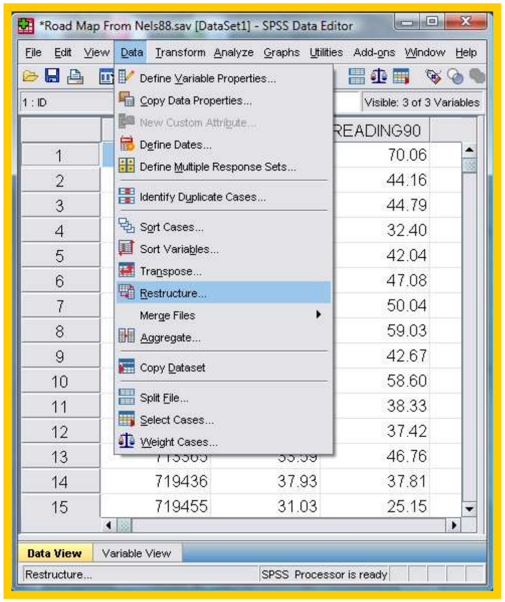


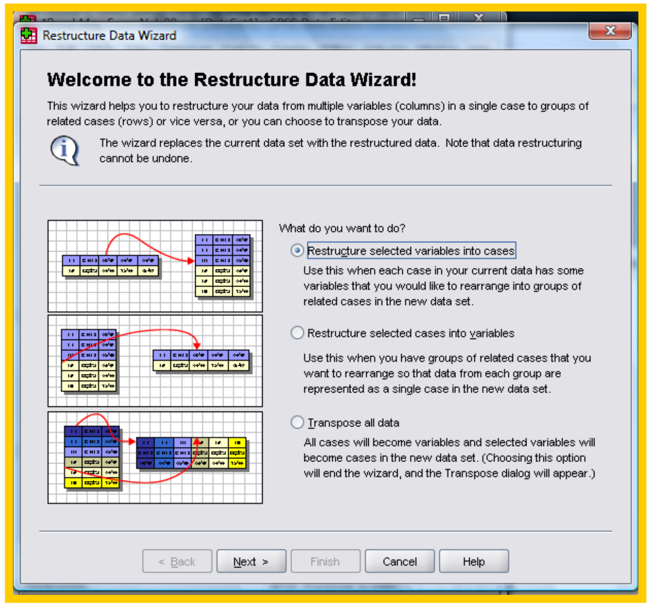
	ID	Index1	READINGL
1	632790	1	51.15
2	632790	2	70.06
3	673838	1	50.90
4	673838	2	44.16
5	673841	1	59.26
6	673841	2	44.79
7	704247	1	28.35
8	704247	2	32.40
9	704858	1	40.88
10	704858	2	42.04
11	708928	1	38.13
12	708928	2	47.08
13	709619	1	44.24
14	709619	2	50.04
15	709658	1	42.87

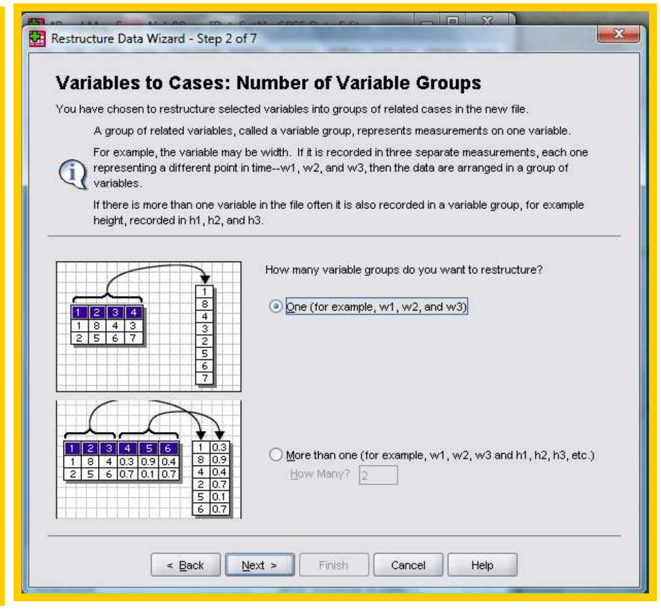
A person-period data set has one time slice per row, but the rows are grouped by an identifying variable and distinguished within the groups by an index variable. No information is lost when converting to person-period data sets.

Note that for most multilevel data, the cluster-observation data set structure is natural. Person-period data sets are the exception. For example, in a mother-child data set, every child will have a mother ID and a child ID, or in a school-student data set, every student will have a school ID and a student ID.

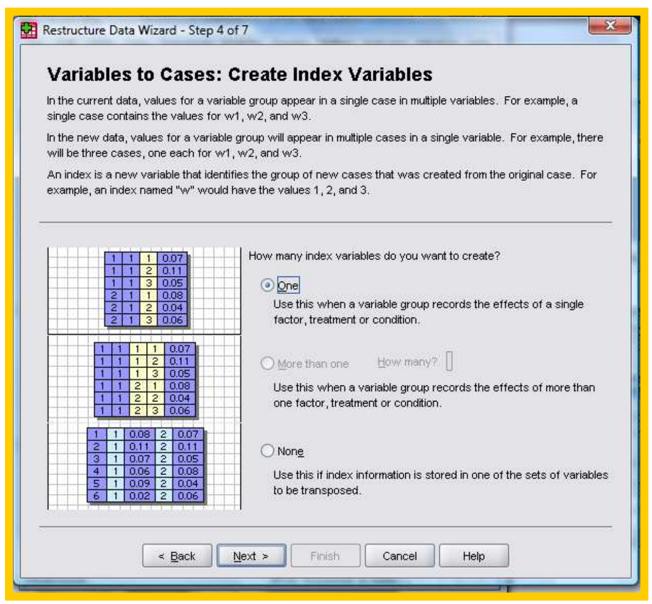
SPSS and Data Set Restructuring

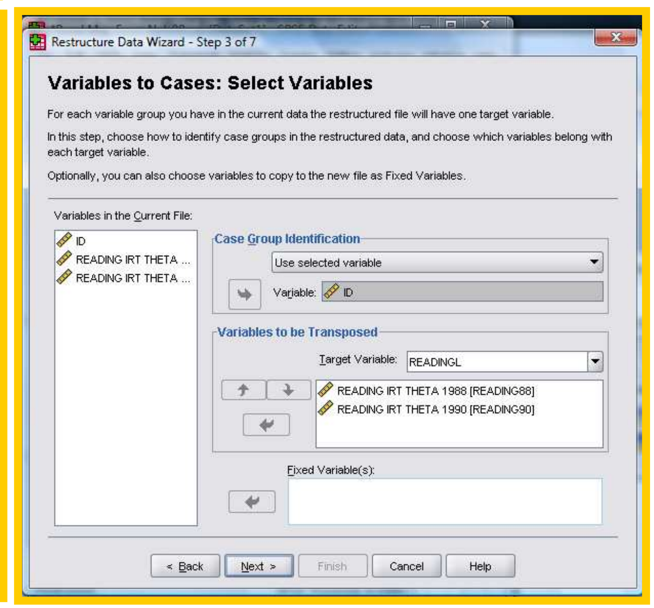






Go to Data > Restructure and SPSS will walk you through all the steps.





Regression Perspective: ε_{ij} and u_i Error Terms (Part I of III)

Now that we have one outcome, we can ask about the mean and variance of THE outcome.

Statistics

READINGL

N	Valid	11856
	Missing	0
Mean		50
Variance		86

Model 0:

Not Quite Right!

$$READINGL = \beta_0 + \varepsilon$$

However, we know that there is a multilevel structure to our data and, consequently, to our outcome. We know that a portion of the variation in scores is attributable to the fact that some students are better readers than other students. We also know that a portion of the variation in scores is attributable to the fact that students improved from the 8th grade to the 10th grade. In other words, we have person-level variation and period-level variation. In still other words, we have student-level variation and score-level variation (where “score” refers to the differing scores for each student depending on wave).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
0	.000a	.000	.000	9.29231

a. Predictors: (Constant)

ANOVAb

Model		Sum of Squares	df	Mean Square	F	Sig.
0	Regression	0	0	0	0	1.00a
	Residual	1023643.544	11855	86.347		
	Total	1023643.544	11855			

a. Predictors: (Constant)

b. Dependent Variable: READINGL

Model 0:

That’s Right!

$$READINGL_{ij} = \beta_0 + \varepsilon_{ij} + u_i$$

Coefficientsa

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
0	(Constant)	50.070	.08534		586.703	.000	49.902	50.237

a. Dependent Variable: READINGL

ε_{ij} represents the residual for the *j*th score of the *i*th student over and above u_i, which represents the residual for the *i*th student.

Regression Perspective: ε_{ij} and u_i Error Terms (Part II of III)

Statistics

READINGL

N	Valid	11856
	Missing	0
	Mean	50
	Variance	86

86.4 = 24.7 + 61.7

Command SPSS to fit an intercept-only model (i.e., unconditional model) that takes into consideration the multilevel structure of the data.

Specify your outcome variable.

Specify your clustering variable.

```
MIXED READINGL
 /PRINT=SOLUTION
 /RANDOM INTERCEPT | SUBJECT(ID).
```

We are now touching on the distinction between random effects and fixed effects in the general linear model. Up until now, we have only dealt with fixed effects models. Now, we are dealing with a mixed model: part fixed, part random. But, let's save a deep discussion of random effects for another day. (We are in deep enough already!)

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	50.069454	.111743	5927.000	448.077	.000	49.850397	50.288511

a. Dependent Variable: READINGL.

Model 0:

$$READINGL_{ij} = \beta_0 + \varepsilon_{ij} + u_i$$

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Residual		24.664534	.453037
Intercept [subject = ID]	Variance	61.687657	1.378449

a. Dependent Variable: READINGL.

Equivalent, "Random Intercepts" Model:

$$READINGL_{ij} = (\beta_0 + u_i) + \varepsilon_{ij}$$

Regression Perspective: ε_{ij} and u_i Error Terms (Part III of III)

The intraclass correlation is the proportion of total variance attributable to the cluster level. When the intraclass correlation is extremely high, all the observations within each cluster are basically the same with respect to the outcome variable. When the intraclass correlation is extremely low, observations within each cluster are clustered together in name only since nothing is tying together their outcome values.

$$\sigma_u^2 = 61.7 = \text{Student-Level Variance (I.e., Between-Student Variance)}$$

$$\sigma_\varepsilon^2 = 24.7 = \text{Score-Level Variance (I.e., Within-Student Variance)}$$

$$\frac{\sigma_u^2}{\sigma_e^2 + \sigma_u^2} = \frac{61.7}{24.7 + 61.7} = .71 = \text{Intraclass Correlation}$$

Whereas the *t*-test and ANOVA uses the Pearson correlation, regression uses the intraclass correlation to account for the non-independence (i.e., clustering) of observations.

Model 0:

$$READINGL_{ij} = \beta_0 + \varepsilon_{ij} + u_i$$

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Residual		24.664534	.453037
Intercept [subject = ID]	Variance	61.687657	1.378449

a. Dependent Variable: READINGL.

??? Guessing Intraclass Correlations ???

In studies of students nested within schools, what is the intraclass correlation? The answer is going to depend on our outcome. Reading scores? Emotional disorders? Community service? Self esteem? Locus of control? For giggles, suppose that our outcome has to do with school clothing, and our data include students clustered within schools. Below are two school-clothing studies, each with its own data set. Which of the two data sets will have the higher intraclass correlation?

Study 1

Study 2

Regression Perspective: Fitting Our Final Model

Model 1:

Command SPSS to fit a model (i.e., conditional model) that takes into consideration the multilevel structure of the data.

$$READINGL_{ij} = \beta_0 + \beta_1 WAVE_{ij} + \varepsilon_{ij} + u_i$$

```
MIXED READINGL WITH WAVE
 /PRINT=SOLUTION
 /FIXED=WAVE
 /RANDOM INTERCEPT | SUBJECT(ID).
```

Specify your predictor variable(s).

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Residual		17.333803	.318413
Intercept [subject = ID]	Variance	65.353023	1.368999

a. Dependent Variable: READINGL.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	48.154558	.118104	7296.199	407.731	.000	47.923040	48.386076
WAVE	3.829793	.076473	5927.000	50.080	.000	3.679878	3.979707

a. Dependent Variable: READINGL.

$$\hat{READINGL} = 48.2 + 3.8WAVE$$

We can use all our MR modeling skills (controlling, interacting, and taxonomizing) to build this simple, one-predictor model into a fully-fledged multiple regression model.

Interpret your fitted multilevel regression model as you would interpret any fitted regression model. But, do so with more confidence because you have not ignored the independence assumption!

You may note that estimated difference between waves 0 and 1 and the associated standard error and t-value is identical to those from the paired samples t-test. We have come back full circle.

Regression Perspective: Presenting Our Final Model

Table 19.1. Fitted multilevel model describing the relationship between a student's IRT scaled reading score and the wave in which the student took the test, where ***WAVE***=0 was the 8th grade (1988) and ***WAVE***=1 was the 10th grade (1990) (n students = 5928, n waves = 2).

	Model	
	M0	**M1**
Intercept	50.1***	48.2***
WAVE		3.8***
σ_u^2	61.7	65.4
Pseudo - R_u^2		0.00
σ_ε^2	24.7	17.3
Pseudo - R_ε^2		0.30

Key: * p<.05; ** p<.01; *** p<.001

On average, in the population, students improve 3.8 points on the IRT scaled reading test from the 8th grade to the 10th grade. Based on a pseudo-R^2 statistic of 0.30, WAVE predicts 30% of the within-student variation in IRT scales reading scores.

We present and interpret our final model just as we would any regression model, except we include our unconditional model (i.e., intercept-only model) as a baseline. From this baseline, we can compare cluster-level variances and observation-level variances.

$$\text{Pseudo-}R_\varepsilon^2 = 1 - \frac{\text{conditional } \sigma_\varepsilon^2}{\text{unconditional } \sigma_\varepsilon^2} = 1 - \frac{17.3}{24.7} = 0.30$$

$$\text{Pseudo-}R_u^2 = 1 - \frac{\text{conditional } \sigma_u^2}{\text{unconditional } \sigma_u^2} = 1 - \frac{65.4}{61.7} = \text{Blech}$$

The pseudo-R^2 statistic is a nice (but sometimes flawed) way to describe the goodness of fit. The true R^2 statistic in the OLS regression to which we are accustomed describes the proportion of variance in the outcome that is predicted by the predictor(s). Now that there are two variances associated with the outcome, we want two R^2 statistics, one for each type of variation—cluster-level variation (i.e., between-cluster variation) and observation-level variation (i.e., within-cluster variation). However, we are no longer doing ordinary least squared regression (OLS). Instead of fitting our model based on the least sum of squares, we are fitting our model based on the least -2 log likelihood. Therefore, our R^2 statistic is not a true R^2 statistic but a pseudo-R^2 statistic. In a multilevel model, the pseudo-R^2 statistic is prone to breaking down when we include only cluster-level variables or only observation-level variables.

Regression: Exploratory Data Analysis and Assumption Checking

Hitherto, we have neglect the crucial book ends to regression modeling, exploratory data analysis and assumption checking. We can (and should!) use all the tools that we have learned in these regards, but twice over. Because we have two levels (the cluster-level and the observation-level), we want to explore each level and check the residuals association with each level.

Exploratory Data Analysis

- SPLASH, DOLMAS and ABORT for the cluster-level data. Use the mean observation for each cluster.
- SPLASH, DOLMAS and ABORT for the observation-level data. Use each observation, but subtract away the mean observation from its respective cluster.

***Obtaining mean observations for each cluster.**

***Obtaining observations minus cluster mean.**

This is not finished, but for now, you can find the SPSS code in this article:

http://www.upa.pdx.edu/IOA/newsom/mlrclass/ho_centering%20in%20SPSS.pdf

Assumption Checking

- Examine RVF plots using residuals from the cluster level, u_i.
- Examine RVF plots using residuals from the observation level, ε_{ij}.

***Obtaining cluster-level residuals.**

***Obtaining observation-level residuals.**

This is not finished, but for now, you can find the SPSS code in this article:

http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/reviewspss.pdf

t-tests, ANOVAs, Regressions, Oh My!

Question: If *t*-tests, ANOVAs and regressions yield identical results, why ever choose the complex ANOVA or the even more complex regression over the simple t-test?

Answer: FLEXIBILITY

	t-test	ANOVA	Regression
Once Repeated Measures	Yes	Yes	Yes
Multiply Repeated Measures	No	Yes	Yes
Categorical Predictors (with or without interactions)	No	Yes	Yes
Continuous Predictors (without interactions)	No	Yes	Yes
Continuous Predictors (with interactions)	No	No	Yes
Any Cluster-Observation Data (e.g., students within schools or children within mothers)	No	No	Yes!

There are an infinite number of error structures that we can specify in multilevel regression modeling, and we touched on the most basic. Consider scores nested within students nested within various teachers nested within schools.

Multilevel regression modeling is known by many names, including “mixed modeling,” “nested modeling” and “hierarchical linear modeling (HLM).” Unfortunately, “HLM” is not only the acronym for hierarchical linear modeling, but it is also the name of proprietary software. You can use HLM (the proprietary software) to do HLM, but you can do HLM in most software packages, including SPSS.