Research Design - - Topic 9

Fundamentals of Bivariate Regression and Correlation

- Bivariate regression (b) - - defining formulae
- Bivariate correlation (r) - - defining formulae
- Test of significance for regression
- An example showing the distinction between b and r
- Interpretations of correlation
- Three limited truths
- Factors that influence the magnitude of r
- Special cases of the Pearson correlation
- Tests of significance
- Correlations with simple aggregates

1

Bivariate Regression and Correlation

Bivariate regression refers to an equation that relates a dependent variable to an independent variable, or a criterion to a predictor. The fundamental equation in raw score form is:

$$Y' = a + b_{yx}X$$

with a and b determined such that $\Sigma(Y-Y')^2$ = a minimum.

and $$a = \overline{Y} - b_{yx}\overline{X}$$

$$b_{yx} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}$$

The formula in standard score form is:

$$Z'_Y = r_{XY}Z'_X$$

where r is as defined on the next slide

2

Bivariate correlation refers to covariation between two variables, X and Y. The most common measure is the Pearson product-moment correlation coefficient defined as:

$$r_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{nS_{b_X}S_{b_Y}} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{(n-1)S_{u_X}S_{u_Y}}$$

$$= \frac{\sum Z_X Z_Y}{n} = \frac{\sum Z_X Z_Y}{n-1}$$

using biased (S_b) and unbiased (S_u) estimates of the standard deviations respectively.

Or alternatives:

$$\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}} = \frac{N\sum XY - \sum X\sum Y}{\sqrt{(N\sum X^2-(\sum X)^2)(N\sum Y^2-(\sum Y)^2)}}$$

3

Given Y=Y' + (Y-Y'), we can compute:

$$\sum(Y-\overline{Y})^2 = \sum(Y'-\overline{Y})^2 + \sum(Y-Y')^2$$

SS_{TOTAL} = $SS_{REGRESSION}$ + $SS_{RESIDUAL}$

And with some algebra, we can construct the following summary table

Source	df	Sums of Squares	
Regression	1	r^2SS_{TOTAL}	$F = \frac{r^2SS_{TOTAL}}{\frac{SS_{TOTAL}(1-r^2)}{n-2}}$
Residual	$n-2$	$SS_{TOTAL}(1-r^2)$	
Total	$n-1$		$= \frac{r^2}{(1-r^2)/(n-2)}$

4

Consider the sample data set:

	X	Y	Z_x	Z_y
	3	3	-1.50	-1.50
	4	5	-.75	-.50
	4	5	-.75	-.50
	4	3	-.75	-1.50
	5	7	0	.50
	5	6	0	0
	5	7	0	.50
	6	9	.75	1.50
	7	7	1.50	.50
	7	8	1.50	1.00
Mean	5.0	6.0	.00	.00
S_u	1.33	2.00	1.00	1.00

5

Computing Regression Coefficients and Correlation

$$b_{yx} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} = \frac{20}{16} = 1.25$$

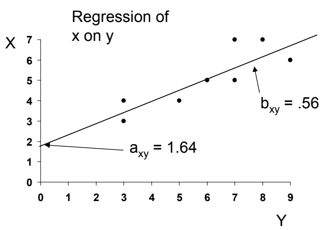
$$a_{yx} = \overline{Y} - b_{yx}\overline{X} = 6.0 - (1.25)(5.0) = -.25$$

$$b_{xy} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (Y-\overline{Y})^2} = \frac{20}{36} = .56$$

$$a_{xy} = \overline{X} - b_{xy}\overline{Y} = 5.0 - (.56)(6.0) = 1.64$$

$$r_{XY} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{(n-1)S_{u_X}S_{u_Y}} = \frac{20}{9(1.33)(2.00)} = .84$$

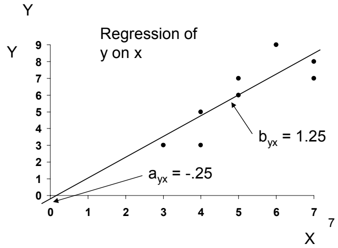
6



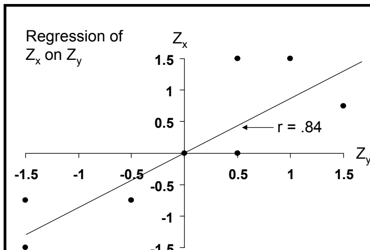
The two regression plots in raw score form

In each case:
a = the **intercept** - - the value of the dependent variable when the independent variable is 0.

b = the **slope** of the regression line (the rise over the run).

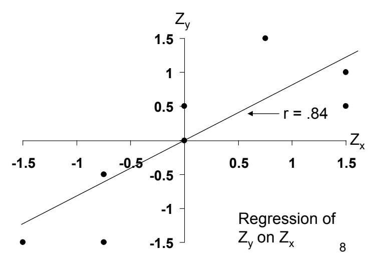


7



The two regression plots in standard score form

Note. The **slope** of the regression line in standard score form is the same in each case and is equal to the correlation coefficient (.84). The **intercept** is 0.



8

Different Interpretations of Correlation

1. Correlation is a measure of the linear relation between y and y':

$$r_{yy'} = \frac{\Sigma(y-\bar{y})(y'-\bar{y})}{\sqrt{\Sigma(y-\bar{y})^2\,\Sigma(y'-\bar{y})^2}}$$

where: $y' = a + bx$

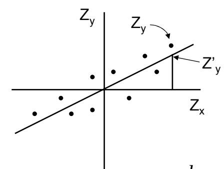
and $a = \bar{y} - b\bar{x}$

$\therefore \quad y' = \bar{y} + b(x - \bar{x})$

$$= \frac{\Sigma(y-\bar{y})b(x-\bar{x})}{\sqrt{\Sigma(y-\bar{y})^2 b^2\,\Sigma(x-\bar{x})^2}} \quad = \quad \frac{\Sigma(y-\bar{y})(x-\bar{x})}{\sqrt{\Sigma(y-\bar{y})^2\,\Sigma(x-\bar{x})^2}} \quad = \quad r_{xy}$$

9

2. Correlation is a measure of the slope of the regression line in standard score form:



r = slope = $\frac{Z'_y}{Z_x} \quad \therefore Z'_y = rZ_x$

Best fit line $\Sigma(Z_y - Z'_y)^2 =$ minimum

$$\Sigma(Z_y - rZ_x)^2 = \Sigma Z_y^2 + r^2\Sigma Z_x^2 - 2r\Sigma Z_x Z_y$$

$$\frac{d}{dr}\left(\Sigma Z_y^2 + r^2\Sigma Z_x^2 - 2r\Sigma Z_x Z_y\right) = 0$$

$$0 + 2r\Sigma Z_x^2 - 2\Sigma Z_x Z_y = 0$$

Note:

$\frac{\Sigma Z_x^2}{n} = S_{Z_x}^2 = 1$

$$r = \frac{\Sigma Z_x Z_y}{\Sigma Z_x^2} = \frac{\Sigma Z_x Z_y}{n} = r_{xy}$$

10

3. Correlation is a measure of the accuracy of predicting y given x:

Given $y = y' + (y - y')$

$S_y^2 = S_{y'}^2 + S_{y-y'}^2$ where y' and (y – y') are independent

$\therefore S_{y'}^2 = S_y^2 - S_{y-y'}^2$

Defining $r_{xy}^2 = \frac{S_{y'}^2}{S_y^2} = \frac{S_y^2 - S_{y-y'}^2}{S_y^2}$

$$r_{xy}^2 = 1 - \frac{S_{y-y'}^2}{S_y^2}$$

$$\therefore r_{xy} = \pm\sqrt{1 - \frac{S_{y-y'}^2}{S_y^2}}$$

11

Three Limited Truths*

1. The Pearson product-moment correlation varies from -1 to +1. True, only under very specific circumstances.

Proof: Given $S_{Z_x}^2 = \frac{\Sigma Z_x^2}{N} = 1$

and $r = \frac{\Sigma Z_x Z_y}{N}$

r can equal +1, only if $Z_x = Z_y$ and -1, only if $Z_x = -Z_y$

Thus, for this to be true, the standardized distributions of x and y must be:

a) Identical
b) Symmetrical (not necessarily normal)

* Adapted from Gardner (2000).

12

2. Given a large enough sample size, the correlation will always be significant. True, only because of artifacts.

Proof: Given $X = T_X + E_{XR} + E_{XM}$ $\quad Y = T_Y + E_{YR} + E_{YM}$

(i.e., the measures of X and Y consist of true scores (T_X & T_Y), random error (E_{XR} and E_{YR}) and measurement error (E_{XM} & E_{YM})).

Given: $\rho_{T_X T_Y} = 0,$

it is possible that $\rho_{XY} \neq 0.$

because the correlations

$$\rho_{T_X E_{YM}},\ \rho_{T_Y E_{XM}} \text{ and } \rho_{E_{XM} E_{YM}} \text{ are not } 0$$

Thus, even with two variables that are truly independent, the correlation between measures of those variables may not be 0, and given a large enough sample size it may be significant.

13

3. Correlation does not mean causation. This is not a limitation of the statistic, but rather the nature of the underlying design.

Consider an experiment on the effects of the amount of alcohol consumed in the afternoon and number of hours slept that night. This study could be run in controlled conditions with careful attention to detail, etc.

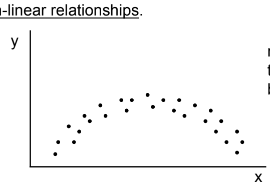
The correlation between the two could be considered an index of the linear effects of alcohol on hours slept (and an indication of causality) if the amount consumed was randomly determined and administered by the experimenter.

The correlation between the two would simply be an index of the covariation between the two if the amount consumed was not determined randomly. The regression equation would describe the nature of the linear relationship.

14

Factors That Affect the Pearson Product Moment Correlation Coefficient

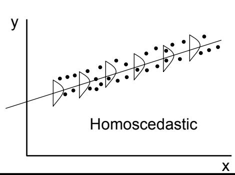
1. Non-linear relationships.

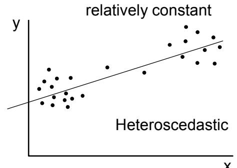


r_{xy} will approach 0, even though there is a non-linear relationship between the variables.

2. Heteroscedasticity.

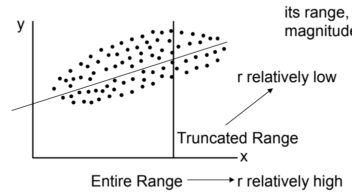
An assumption underlying r_{xy} is homoscedasticity – viz., that the variation around the regression line is relatively constant





15

3. Restriction in Range.



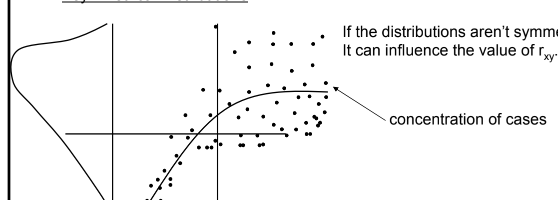
If selection on one variable has restricted its range, this could influence the magnitude of r_{xy}. To correct for this, calculate

$$r_{cor} = \frac{r_{xy}\, S_x / s_x}{\sqrt{1 - r_{xy}^2 + r_{xy}^2\, S_x / s_x}}$$

where S_x = stand. dev. uncurtailed

s_x = stand. dev. curtailed

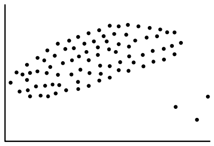
4. Asymmetrical Distributions.



If the distributions aren't symmetrical, It can influence the value of r_{xy}.

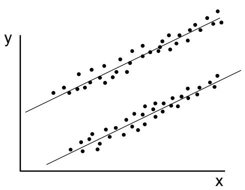
16

5. The Presence of Outliers.



A few very extreme scores can influence the correlation.

6. A Confounding Variable.



Mixing samples can sometimes eliminate meaningful relationships. Assume X = IQ, Y = grades. Two classes may be similar on IQ, but differences in level of teacher grading might obscure relation between grades and IQ if the samples are merged.

7. Reliability.

The correlation between two measured variables underestimates the correlation between the true variables in the population as the reliability of the measures decreases. The correction for such attenuation is:

$$r_{T_x T_y} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

where r_{xx} and r_{yy} are the reliabilities of the measured variables.:

Special Cases of the Pearson Product Moment Correlation

Spearman Rank Order = ρ_s

Correlation between two variables, ranked from 1 to N.

$$\sum x = \sum y = \frac{N(N+1)}{2} \qquad \sum x^2 = \sum y^2 = \frac{N(N+1)(2N+1)}{6}$$

Given: $d = x - y$

$$\sum d^2 = \sum x^2 + \sum y^2 - 2\sum xy$$

$$\therefore \sum xy = \frac{\sum x^2 + \sum y^2 - \sum d^2}{2}$$

$$\therefore r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} = 1 - \frac{6\sum d^2}{N(N^2-1)} = \rho_s$$

Point Biserial = r_{pb}

Correlation between a dichotomous and continuous variable.

x

0	1
y_{1i}	y_{2i}
y_{1i}	y_{2i}
•	•
•	•
•	•
•	•
$\bar{y}_1$	$\bar{y}_2$
n_1	n_2

$$\sum x = \sum x^2 = n_2 \qquad \sum xy = n_2 \bar{y}_2$$

$$N = n_1 + n_2 \qquad p = n_1/N \qquad q = n_2/N$$

$$\therefore r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} = \frac{(\bar{y}_2 - \bar{y}_1)\sqrt{pq}}{S_y} = r_{pb}$$

Phi Coefficient = Φ

Correlation between two dichotomous variables.

y \ x	0	1	
1	b	a	a + b
0	d	c	c + d
	b + d	a + c	N = a + b + c +d

$$\sum x = \sum x^2 = a + c$$
$$\sum y = \sum y^2 = a + b$$
$$\sum xy = a$$

$$\therefore r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)(a+c)}} = \phi$$

$$and\ \phi = \sqrt{\frac{x^2}{N}} \quad for\ 2\ x\ 2\ tables$$

Effect Strength and Power

Cohen's (1988) definitions:

Small ρ=.10. "many relationships pursued in "soft" behavioral sciences are of this magnitude" (p. 79).

Medium ρ=.30. "this degree of relationship would be perceptible to the naked eye of a reasonably sensitive observer" (p. 80).

Large ρ=.50. "around the upper end of the range of (nonreliability) r's one encounters in those fields of behavioral science which use them extensively" (p.80).

Power can be calculated using Cohen (1988) or G*Power3. Note that G*Power3 has three routines that can be used for this purpose, one with t, one with F, and one with Exact tests.

Testing the significance of a single bivariate correlation coefficient

1. *Ho:* ρ = 0.

$$F = \frac{r^2}{(1-r^2)/(N-2)}$$

$$@\ df_1 = 1; \quad df_2 = N - 2.$$

or its equivalent:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

$$@\ df = N - 2.$$

2. $Ho:\ \rho = 0. \quad (For\ large\ N).$

$$Z = r\sqrt{N-1}$$

Testing the significance of a single multiple correlation coefficient

3. *Ho:* ρ = 0.

$$F = \frac{R^2/p}{(1-R^2)/(N-p-1)}$$

$$@\ df_1 = p; \quad df_2 = N - p - 1$$

Testing the difference between two correlation coefficients

4. $Ho: \ \rho_1 = \rho_2$ *for independent samples.*

Fisher's Z
$$Z_{r1} = \frac{1}{2}\log_e \frac{(1+r_1)}{(1-r_1)}$$

$$Z_{r2} = \frac{1}{2}\log_e \frac{(1+r_2)}{(1-r_2)}$$

and
$$Z = \frac{Z_{r1} - Z_{r2}}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$$

25

Comparing two correlations from the same sample (with a common variable)

5. $Ho: \ \rho_{12} = \rho_{13}$ *for correlated correlations.*

1. Test proposed by Dunn and Clark (1969).

$$Z = \frac{(r_{12} - r_{13})\sqrt{N}}{\sqrt{(1-r_{12}{}^2)^2 + (1-r_{13}{}^2)^2 - 2r_{23}^3 - (2r_{23} - r_{12}r_{13})\,(1 - r_{12}{}^2 - r_{13}{}^2 - r_{23}{}^2)}}$$

2. Test proposed by Meng, Rosenthal & Rubin (1992).

$$Z = (Z_{r1} - Z_{r2})\sqrt{\frac{N-3}{2(1-r_{23})h}}$$

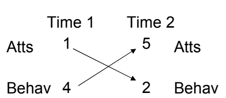
where each: $Z_r = \frac{1}{2}\log_e \frac{(1+r)}{(1-r)}$

and $f = \dfrac{1 - r_{23}}{2(1 - (r_{12}^2 + r_{13}^2)/2)}$ $\qquad h = \dfrac{1 - f\,(r_{12}^2 + r_{13}^2)/2}{1 - (r_{12}^2 + r_{13}^2)/2}$

26

Comparing two correlations from the same sample (with different variables) (Cross Lagged Panel Analysis)

6. $Ho: \ \rho_{12} = \rho_{45}$ *for correlated correlations.*



$$Z = \frac{(r_{12} - r_{45})\sqrt{N}}{\sqrt{(1-r_{12}{}^2)^2 + (1-r_{45}{}^2)^2 - r_{12}r_{45}(r_{14}{}^2 + r_{15}{}^2 + r_{24}{}^2 + r_{25}{}^2) - 2(r_{14}r_{25} + r_{15}r_{24}) + C}}$$

where:

$$C = 2(r_{12}r_{14}r_{15} + r_{12}r_{24}r_{25} + r_{14}r_{24}r_{45} + r_{15}r_{25}r_{45})$$

27

Testing the Significance of an average correlation

7. $Ho: \ \rho_{av} = 0$

$$Z_{AV} = \frac{(n_1 - 3)Z_{r1} + (n_2 - 3)Z_{r2} + \ldots + (n_k - 3)Z_{rk}}{(n_1 - 3) + (n_2 - 3) + \cdots + (n_k - 3)}$$

where:

$$Z_r = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

then:

$$Z = Z_{AV}\sqrt{((n_1 - 3) + (n_2 - 3) + \cdots + (n_k - 3))}$$

28

Testing the significance of a partial correlation

8. $Ho: \quad \rho_{12.3} = 0$

$$t = \frac{r_{12.3}}{\sqrt{(1 - r_{12.3}^{\ 2})/(N-3)}} \qquad df = N - 3.$$

Testing the significance of a semipartial (part) correlation

9. $Ho: \quad \rho_{1(2.3)} = 0$

$$F = \frac{(N-3)\, r_{1(2.3)}^{2}}{1 - R_{1.23}^{2}} \qquad df = 1, \quad N-3$$

Note. These two statistics yield identical results, except that F = t² (both at N-3 df).

Comparing two bivariate regression coefficients

10. $Ho: \quad b_1 = b_2 \ in\ the\ population$

$$t = \frac{b_1 - b_2}{S_{Db}}$$

where:

$$S_{Db} = C\sqrt{\frac{n_1 S_{y1}^{\ 2}(1 - r_1^{\ 2}) + n_2 S_{y2}^{\ 2}(1 - r_2^{\ 2})}{(n_1 + n_2 - 4)}}$$

and:

$$C = \sqrt{\frac{1}{n_1 s_{x1}^{\ 2}} + \frac{1}{n_2 s_{x2}^{\ 2}}}$$

and df = $n_1 + n_2 - 4$.

Correlations Involving Aggregates

Raw Data

X_1	X_2	X_3	Y	T_X	T_Z
6	10	32	100	48	-.51
8	9	25	97	42	-2.34
10	13	31	103	54	2.37
9	13	29	106	51	1.06
10	15	30	105	55	2.66
7	9	27	92	43	-2.15
6	10	29	85	45	-1.64
11	18	26	106	55	2.73
7	9	24	90	40	-3.28
9	12	30	93	51	1.10

For these data:

$T_Z = Z_{X1} + Z_{X2} + Z_{X3}$

$r_{T_Z y} = .769$

$T_X = X_1 + X_2 + X_3$

$r_{T_X y} = .754$

Correlation Matrix

	Y	X_1	X_2	X_3
Y	1.0000	.7415	.7346	.2675
X_1	.7415	1.0000	.8688	.0259
X_2	.7346	.8688	1.0000	.1742
X_3	.2675	.0259	.1742	1.0000

Aggregated Standard Scores:

$$r_{T_Z y} = \frac{\sum_{j=1}^{m} r_{jy}}{\sqrt{\sum_{j=1}^{m}\sum_{k=1}^{m} r_{jk}}} = \frac{.7415 + .7346 + .2675}{\sqrt{1.000 + .8688 + \cdots + .1742 + 1.000}}$$

$$= \frac{1.7436}{\sqrt{5.1378}} = \frac{1.7436}{2.2667} = .769$$

Covariance Matrix

	Y	X_1	X_2	X_3
Y	55.503	9.778	16.473	5.321
X_1	9.778	3.133	4.629	.122
X_2	16.473	4.629	9.060	1.400
X_3	5.321	.122	1.400	7.129

Aggregated Raw Scores:

$$r_{T_x y} = \frac{\sum_{j=1}^{m} \text{cov}_{jy}}{S_y \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{m} \text{cov}_{jk}}} = \frac{9.778 + 16.473 + 5.321}{\sqrt{55.503}\sqrt{3.133 + 4.629 + \cdots + 1.400 + 7.129}}$$

$$= \frac{31.572}{(7.450)\sqrt{31.624}} = \frac{31.572}{41.895} = .754$$

Correlations Involving Difference Scores

(1) Correlation of Initial Score with the Difference

$$r_{x(y-x)} = \frac{\sum(x-\bar{x})[(y-x)-(\bar{y}-\bar{x})]}{N\, S_x\, S_{y-x}}$$

$$= \frac{M\, r_{xy} - 1}{\sqrt{1 + M^2 - 2Mr_{xy}}}$$

$$where: \quad M = \frac{S_y}{S_x}$$

(2) Correlation of one variable (A) with a Difference (y – x)

$$r_{A(y-x)} = \frac{\sum(A-\bar{A})[(y-x)-(\bar{y}-\bar{x})]}{N\, S_A\, S_{y-x}}$$

$$= \frac{M\, r_{Ay} - r_{Ax}}{\sqrt{1 + M^2 - 2Mr_{xy}}} \qquad where: \quad M = \frac{S_y}{S_x}$$

(3) Correlation between two Difference Scores

$$r_{(B-A)(y-x)} = \frac{\sum[(B-A)-(\bar{B}-\bar{A})][(y-x)-(\bar{y}-\bar{x})]}{N\, S_{B-A}\, S_{y-x}}$$

$$= \frac{M\,(Lr_{By} - r_{Ay}) - (Lr_{Bx} - r_{Ax})}{\sqrt{[L^2 + 1 - 2Lr_{AB}][M^2 + 1 - 2Mr_{xy}]}}$$

$$where: \quad M = \frac{S_y}{S_x} \quad ; \quad L = \frac{S_B}{S_A}$$

References

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (Second Edition).* Hillsdale, NJ: Lawrence Erlbaum.

Dunn, O.J. & Clark, V.A. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64, 36*6-377.

Gardner, R. C. (2000). Correlation, causation, motivation and second language acquisition. *Canadian Psychology, 41,* 10-24.

Gardner, R.C. & Erdle, S. (1984). Aggregating scores: To standardize or not to standardize? *Educational and Psychological Measurement, 44,*813-822.

Gardner, R.C. & Neufeld, R.W.J. (1987). Use of the simple change score in correlational analyses. *Educational and Psychological Measurement, 47,* 849-864.

Meng, X-L, Rosenthal, R. & Rubin, D.B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111, 172*-175.

Tabachnick, B. G. & Fidell, L.S. (2007). *Using Multivariate Statistics (Fifth Edition).* Needham Heights, MA: Allyn & Bacon.