What Is the Impact on Pearson Correlation Coefficients and the R Squared Measure of Fit in Regression When Some of the Respondents Are Not Able to Follow the Rating Scale?

Suzanne L. Slocum*, Charles O. Ochieng**,
& Bruno D. Zumbo*
*The University of British Columbia
**CTB/McGraw-Hill

Paper Presented at the 2003 American Educational Research Association (AERA) annual meeting in Chicago, Illinois

Introduction to the Problem

- As was stated in the first paper in this symposium (Zumbo & Ochieng, 2003) the central issue is that there is an unspoken measurement assumption when we analyse Likert / rating scale data.
- The assumption is that everyone in your population is responding using the same response process – that is, the same thresholds for a given variable.
- Our question is directed toward ordinary least-squares regression and Pearson correlation coefficients when this assumption is not true.

Problem in Context

- Imagine a school/counseling psychology researcher is interested in regressing the response to a depression item -How often have you cried in the last two weeks?- onto three continuous variables:
  - X1= age
  - X2= number of recent visits to the school counselor
  - X3= total score on an emotional sensitivity scale
- The dependent (y) variable is a variable that may range from 3 to 9 points depending on the scale format the researcher is using.

The Problem in Context II

- The key issue is that because the y-variable in our regression is not continuous but rather a Likert format we may be faced with some mis-responding (that is, different thresholds are being used).
- For example, for a crying item it is documented in the literature that men and women respond differently, and that men tend to have higher thresholds on the crying item.
- If we have males and females in our sample, and if the males are using a different response process, what are the effects on the R-squared in the regression described above, and on the correlations among dependent and independent variables?

Purpose

- Investigating the effect of having Likert scale mis-responding on:

1. The R-squared from OLS regression where only Y is a Likert variable.
2. The results of Pearson correlation among two variables when:
    a. both variables are Likert and hence both <u>may</u> have mis-responding,
    b. only one variable is Likert and hence only one variable <u>may</u> have mis-responding.

Past Research

- In the continuous variable case it is known that both the OLS R-squared from regression and the Pearson correlation coefficient are slightly negatively biased (see a recent paper by Zimmerman, Zumbo, & Williams, 2003).
- It is also known in the research literature that both the R-squared from OLS regression and the Pearson correlation are attenuated by Likert variables (Bollen & Barb, 1981; Ochieng 2001) but that this attentuation is reduced with increasing scale points – with little marginal gain after 5 scale points

Research Purpose

- What, to our knowledge, is largely undocumented is what happens to R-squared from an OLS regression and to Pearson correlations when the rating scale response process is different for some sub-group, in our example, between males and females.
- Because Likert / rating scale variables are widely used in research and likewise because OLS regression and Pearson correlations are widely applied to this Likert data, the question we are addressing is of much import to day-to-day data analysts and researchers.

Simulation Methodology: Factors in the Study

- The overall methodology is the one described in the opening paper by Zumbo and Ochieng (2003).

1. Multiple Regression with 3 x-variables
   a. The x-variables are continuous (as in our example above)
   b. The y-variable has values ranging from 3 – 9 scale points and is symmetric.
2. Four different proportions of people who have different thresholds: 0, 10, 20, and 30 percent of mis-responders

- We have a 7 x 4, scale points by percentage misresponders, completely crossed design with one observation per cell of our simulation.

Methodology Continued

- For the situations wherein we have misresponding, we have, by simulation design, an indicator variable of whether that response was from a mis-responder or not. In the case of our example where mis-responding is related to gender of the respondent, we would have the gender indicator variable.
- This allows us to investigate regression modeling without and with the subpopulation indicator.
  a. The former allows us to investigate results of the regression wherein the researchers is unaware of the subpopulation of misresponders and unknowningly performs the regression over the entire population of respondents.
  b. The latter allows us to investigate what occurs if the researcher has an indicator variable that perfectly identifies the sub-population of misresponders.

Methodology Continued

In short:

- What would the results be if the researcher did not know that there is sub-population of mis-responders and hence proceeded unknowingly?
- What would the results be if the researcher had, knowingly or naively, included a sub-population indicator in the model that corresponded with the misresponders? (e.g., gender-based misresponding and gender was included in the regression model)

Methodology Continued

- Reminder: the type of mis-responding involves not being able to differentiate between the top two response scale options and hence responding with the lesser of the two.
- We used the following correlation matrix from Ochieng (2001)

R-squared .753

	Y	X1	X2	X3
Y	1			
X1	.60	1		
X2	.50	.20	1	
X3	.70	.30	.20	1

Analysis Strategy for the simulation study

- Data visualization will be the primary approach of summarizing the results. Statistical models will be fit to the simulation outcome variables as well.
- We know from previous research that the relationship between the number of scale points and R-squared (Ochieng, 2001) is quadratic so we will investigate that first.
- Outcome variables from the simulation include bounded variables such as Pearson's r and the R squared. As standard practice we will examine the statistical model residuals to see if we need to consider a transformation.

Results

- R-squared from OLS Regression, without sub-population indicator

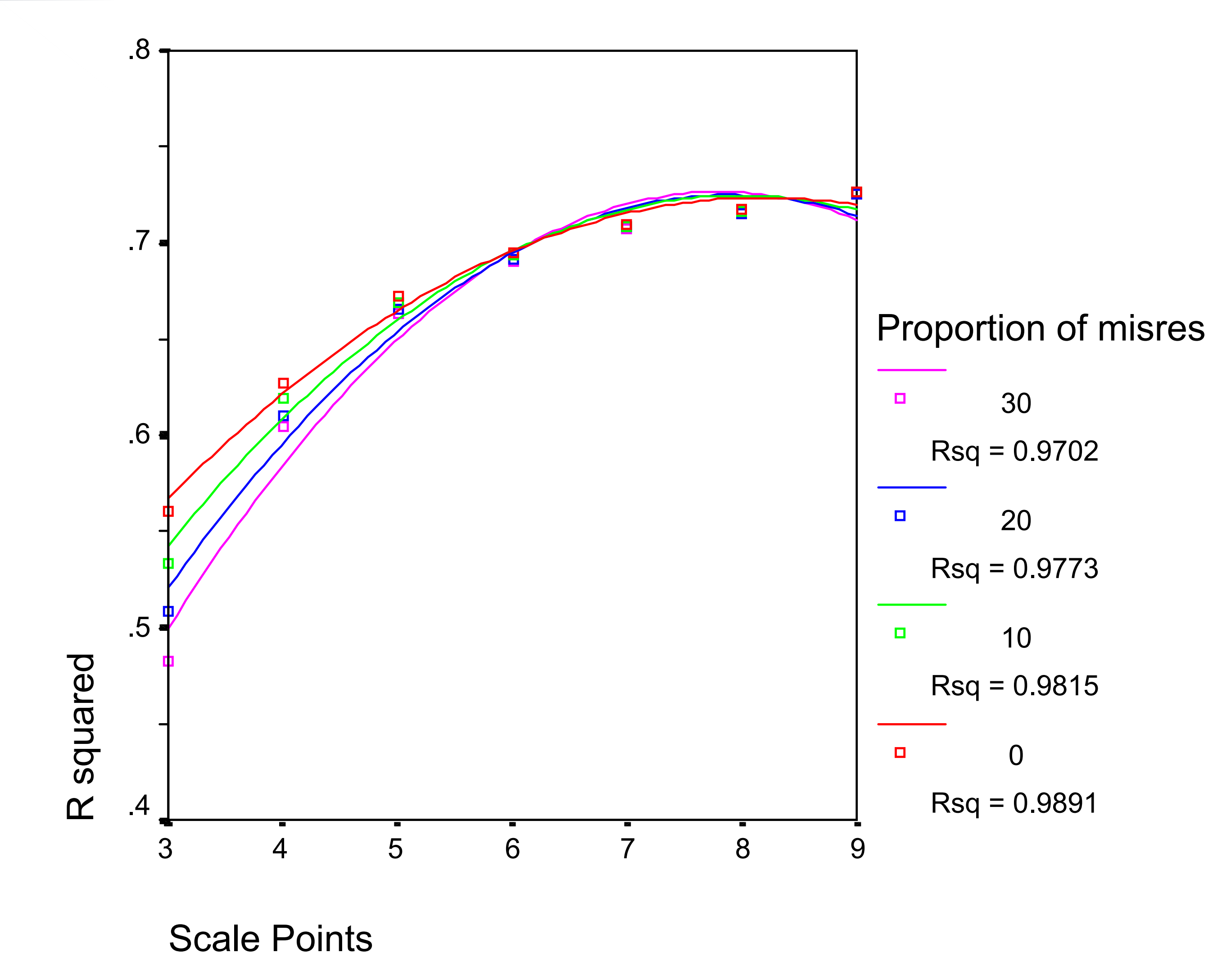
Effect of misresponding

		Percentage of misresponding			
		0	10	20	30
Scale Points	3	.561	.533	.508	.483
	4	.627	.619	.610	.604
	5	.672	.669	.665	.663
	6	.695	.694	.691	.690
	7	.710	.709	.708	.707
	8	.717	.716	.715	.715
	9	.726	.726	.725	.725

Continuous R-squared = 0.753

Results

Graph from table on previous page.



Results

- Clearly from the previous graphs and tables, the results of the simulation appear to change at 4 scale points.
- We investigated this by fitting a linear model to the simulation results separately for less than or equal to 4 scale points, and for 5 or more scale points.
- Each linear model involved: a variable for the number of scale points, a variable for the proportion misresponding, and a third variable that was the interaction of those two variables.
- The results confirm what we saw in the table and graphs.

Results

- For 4 or less scale points, we found:
  - an effect of number of scale points (3 vs 4 points)
  - an effect of the proportion misresponding, and
  - an interaction of scale points by misresponding; that is, the difference between 3 and 4 scale points depends on the proportion of mis-responding.
- For 5 or greater scale points we only found and effect of the number of scale points.

Results

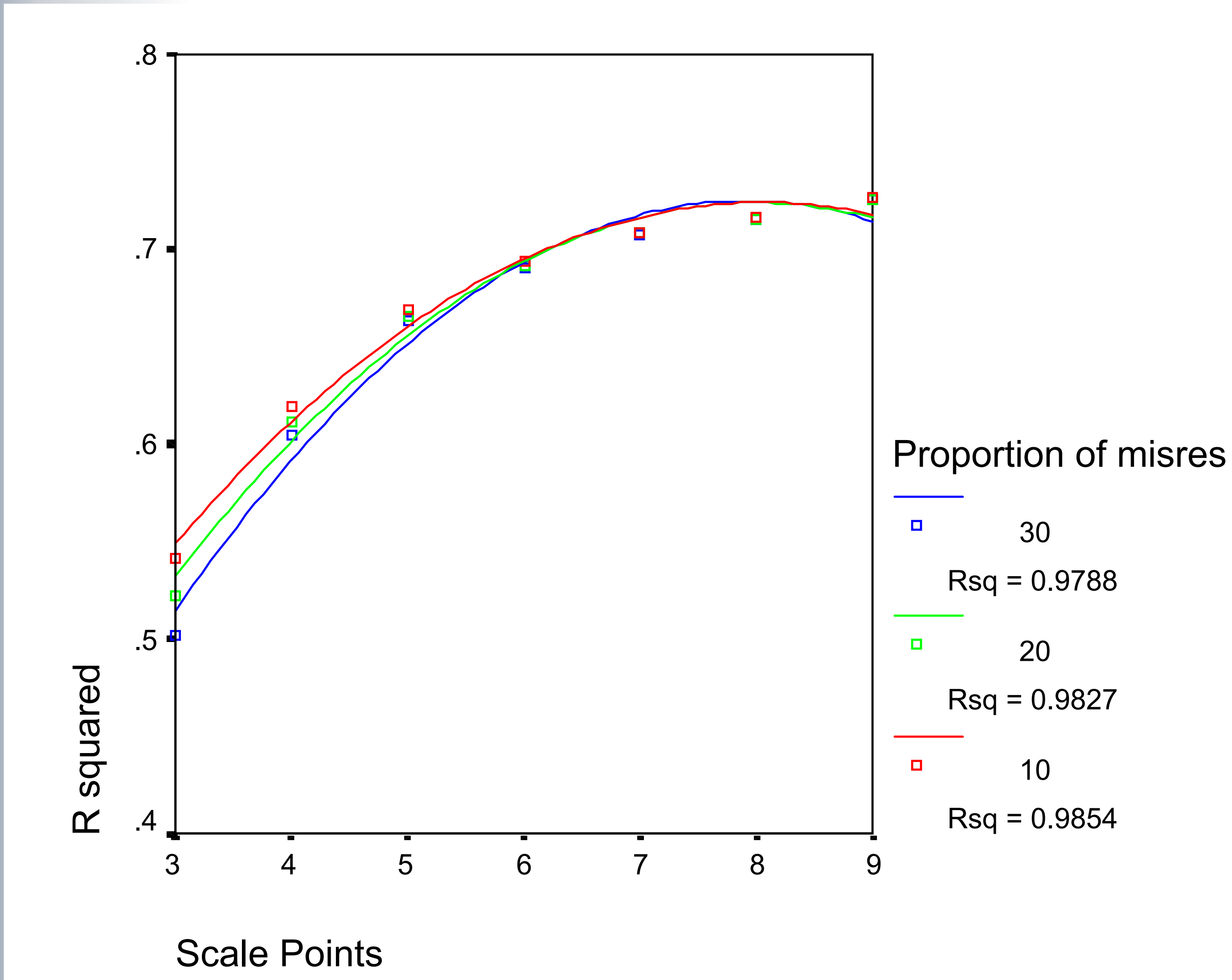
- R-squared, regression, with subpopulation indicator in the model (only when we had misresponding)

		Proportion of misresponding		
		10	20	30
Scale Points	3	.541	.522	.502
	4	.619	.611	.605
	5	.669	.666	.663
	6	.694	.691	.690
	7	.709	.708	.707
	8	.716	.715	.715
	9	.726	.725	.725

The results for this design are the same as not having the indicator variable, but a bit less pronounced.

Results

Graph of R-squared from table on previous page.



Results

- We are now going to turn our attention to the results involving Pearson correlations.
- We investigated the following bivariate correlations from the same simulation data as the regression.
  a. correlation between two Likert variables, misresponding may occur on both
  b. correlation between two Likert variables, misresponding may occur on one
  c. Correlation between a continuous and Likert variables (of course, similar to the R-squared in regression).

Results

Likert to Likert Pearson Correlations

		Proportion of misresponding						
		None	10		20		30	
			both	only y	both	only y	both	only y
scale points	3	.371	.372	.362	.370	.351	.370	.344
	4	.419	.417	.416	.415	.414	.410	.410
	5	.448	.446	.446	.446	.446	.443	.444
	6	.462	.461	.461	.460	.461	.459	.460
	7	.473	.472	.473	.472	.472	.471	.472
	8	.478	.478	.478	.477	.478	.477	.477
	9	.483	.485	.483	.482	.483	.482	.482

Results

Likert to continuous Pearson correlation

		Proportion of misresponding			
		none	10	20	30
Scale Points	3	.432	.421	.411	.401
	4	.458	.455	.451	.449
	5	.474	.474	.472	.472
	6	.482	.482	.481	.481
	7	.488	.487	.487	.487
	8	.491	.491	.490	.490
	9	.493	.492	.492	.492

Conclusions and Future Research

- We want to be clear that we are not suggesting using OLS when the y variable has 3 points, but rather documenting what happens when you use it.
- Overall, we see minimal effect of mis-responding on R-squared and Pearson correlations at 5 to 9 point response scales.
- At the 3 - 4 point response scales there is an effect of misresponding wherein we see that the mis-responding attentuates the results.

Future Research

- In this study we only focused on the large-sample (population analogue) effects of Likert response format and misresponding. We are, in essence, looking at how much we lose by using Likert response format as compared to the continuous unobserved variable in the "population" or some analogue to the population.
- We next need to explore the sample-to-sample variability of the statistics – the standard error and likewise the operating characteristics of Type I error and statistical power. This will involve repeated sampling (perhaps bootstrapping) from the populations to study the Type I error and power.

Future Research

- Potential confounder: We are mimicking a type of mis-responding that actually has a proportionally different effect at the various number of scale points. For example, collapsing the top two scale points is proportionally more for a 3-point scale (2 out of 3) rather than a 9-point scale (2 out of 9).
- This potential confounder may actually be the effect that we are seeing of the 3 and 4 scale points. The “confounding”, however, also reflects a plausible real mis-responding so it is worth considering.

Future Research

- Mis-responding type: we are investigating a narrow type of mis-responding - merging the last two scale points. There are countless misresponding types, and there is a need to further investigate the effects of mis-responding under the various mis-responding conditions.

References

- Bollen, K.A., & Barb, K. (1981). Pearson's r and coarsely categorized measures. American Sociological Review, 46, 232-239.
- Ochieng, C.O. (2001). *Implications of using Likert data in multiple regression analysis*. Unpublished Doctoral Dissertation, University of British Columbia.
- Zumbo, B.D. & Ochieng, C.O. (2002). The Effects of Various Configurations of Likert, Ordered Categorical, or Rating Scale Data on Ordinal Logistic Regression Pseudo R squared Measure of Fit: The case of cummulative logit model. Annual Meeting of the American Educational Research Association (AERA), New Orleans, Lousiana.