Statistical and Graphical Modeling to Investigate Differential Item Functioning for Rating Scale and Likert Item Formats

Suzanne L. Slocum

Michaela N. Gelin

Bruno D. Zumbo

University of British Columbia

Citation:
Slocum, S. L., Gelin, M. N., & Zumbo, B. D. (2004). *Statistical and Graphical Modeling to Investigate Differential Item Functioning for Rating Scale and Likert Item Formats*. Vancouver: Edgeworth Laboratory, University of British Columbia.

Send Correspondence to:
Send Correspondence to: Correspondence regarding this paper can be addressed to Professor Bruno Zumbo, Department of Educational and Counselling Psychology, and Special Education, Measurement Evaluation and Research Methodology, University of British Columbia, 2125 Main Mall, Vancouver, BC, Canada, V6T 1Z4. Electronic mail may be sent to bruno.zumbo@ubc.ca

Date: September 4, 2004

Abstract

Differential item functioning (DIF) is a threat to the inferences made from test scores. DIF is a necessary but not sufficient condition for item bias. That is, if we do not find DIF then we do not have item bias, but if we do find DIF then it may be item bias or it may be item impact. Therefore, it is important so that test score inferences are not compromised that DIF analyses be performed. The definition and purpose of DIF are described in this paper in the context of an example. Furthermore, this paper describes a statistical modeling approach for detecting DIF and introduces a graphical display for studying DIF.

Statistical and Graphical Modeling to Investigate Differential Item Functioning for Rating Scale and Likert Item Formats

Differential item functioning (DIF) and item bias are becoming increasingly important to use when constructing new measures or validating inferences made from test scores (Zumbo & Hubley, 2003). DIF and item bias are threats to the validity of test score inferences. How is DIF a threat? If one does not investigate DIF, and hence only looks at the (unmatched) item performance or only scale-level performance (Zumbo, 2003), one has no way of knowing whether or not the observed differences between groups are due to measurement artifact (Zumbo, 1999).

Without DIF analyses, the differences may be real differences (termed "impact" in the measurement literature) wherein the examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item. On the other hand, the differences may reflect differential responses to items (differing probabilities of success on an item) after matching on the underlying ability that the item is intended to measure (Zumbo, 1999). Item bias occurs when examinees from one group are less likely to answer an item correctly than examinees of another group because of characteristics of the test item or the measurement situation that are not relevant to the testing purpose. DIF is necessary, but not sufficient, for item bias. Once one flags an item as DIF, the item needs to be studied further by content specialists for item bias. That is, if we do not find DIF then we do not have item bias, but if we do find DIF then it may be item bias or it may be item impact, depending on the conclusions of content specialists. One can rule-out item bias but one cannot prove item bias with DIF statistical analyses alone. This is reminiscent of the limitations to the causal inferences one can make in any use of statistical modeling and hypothesis testing.

Having made the case for the importance of DIF analyses, the purpose of this paper is to describe a statistical modeling approach and introduce a graphical display for studying DIF for rating scale or Likert-type items. Recognizing that it more effective to illustrate a statistical technique in the context of an example rather than describe in abstract terms, the remainder of this paper will use the Center of Epidemiology Scale of Depression (CES-D) to motivate the presentation.

Differential Item Functioning of the CES-D

The measurement of depression in the general population is increasingly more common in educational, psychological, and epidemiological research. One of the most widely used instruments is the Center of Epidemiology Scale of Depression (CES-D) (Radloff, 1977).

The CES-D is comprised of 20 items that ask respondents to rate, on a 4-point Likert scale, the frequency in which they experience depressive symptoms. A total score is computed by summing the 20 item responses. The CES-D instructions and items 17 and 20 of the scale are shown in Figure 1. We will be using items 17 and 20 to illustrate the DIF methods.

Insert Figure 1 about here

Participants

The data are from a community-based survey conducted in Northern British Columbia as part of the activities of the Institute for Social Research and Evaluation at the University of Northern British Columbia. The participants are 724 adults between the ages of 17 and 92 years (357 females; 367 males). Males in the study are statistically significantly older (female: 44 years old; male: 47 years old, $t(722)=3.1$, $p=0.002$, eta-squared=0.013), although the effect size is not very large. Missing values for item responses were imputed by using an EM algorithm in the software PRELIS.

How does DIF apply in the context of the CES-D? An example would be gender DIF. It is widely reported in the literature that there are gender differences in the CES-D total scores, which leads us to the question: Are these differences due to measurement artifact? By measurement artifact we mean that the difference is due to the way that depression is being measured. That is, men and women are responding differently to one or more of the items, and this difference in item responses is an *artifact* of the measurement process – item bias.

Measurement Artifact

There are several important methodological points that need to be highlighted when performing DIF analyses. The first is that the focus of the DIF analysis is to study each item individually. And the second is that one needs to perform a statistical matching to make the comparison between genders meaningful. DIF would mean that the typical item responses are different for males and females after matching on total scores; that is taking into account depressive symptoms. See Zumbo and Hubley (2003) for details on matching and the variety of DIF models available.

How do we test for DIF?

In this paper, we will focus on two classes of approaches: the first is the statistical modeling for DIF and the second is a graphical display. When generating a statistical model for DIF, as Zumbo and Hubley (2003) state, many of the DIF techniques are akin to ANCOVA, wherein we statistically condition (match) on a variable of interest and then investigate group differences. Over and above what we are trying to measure, is there a group difference for each item? If so, then there is DIF.

There are several types of regression modeling one can utilize when investigating potential DIF. One can either use logistic regression, ordinal logistic regression, or ordinary regression depending on the dependent variable in the model and the assumptions one is willing to make about the error term. That is, logistic regression is used for binary item scores, ordinal logistic regression is used for rating scale or Likert items, and ordinary least-squares regression is used for continuous or Likert items that have many scale points (e.g., more than 6 scale points). In our example with the CES-D, we used ordinal logistic regression because of the 4-point Likert scale item response format.

Statistical Modeling

DIF methods allow one to judge whether items are functioning the same in various groups. One needs to match these groups on the ability of interest prior to examining whether there is a group effect. One is interested in stating the probability model in studying main effects of group differences (uniform DIF) and interaction of group by ability (non-uniform DIF).

In order to conduct these analyses we used SPSS because it is widely available and used in the social and health sciences. See Appendix A of this paper for SPSS syntax

and drop-down windows for a demonstration of how to conduct ordinal logistic regression. Likewise, see Zumbo (1999) for a detailed description.

As described by Zumbo (1999), the ordinal logistic regression (OLR) procedure uses the item response as the dependent variable, with the grouping variable (denoted as GRP), total scale score for each examinee (denoted as TOTAL) and a group by total interaction as independent variables. This can be expressed as a linear regression of predictor variables on a latent continuously distributed random variable, y*.

The ordinal logistic regression equation is

$$y^* = b_0 + b_1 TOTAL + b_2 GRP + b_3 TOTAL * GRP_i$$

Zumbo's OLR approach provides a test of DIF that measures the effect of group and the interaction, over-and-above the total scale score. This DIF method has a natural hierarchy of entering variables into the model in which the conditioning variable (i.e. the total score) is entered first. Next, the grouping variable (e.g., gender) is entered. This step measures the effect of the grouping variable while holding constant the effect of the conditioning variable (uniform DIF). Finally, the interaction term (e.g., TOTAL*GENDER) is entered into the equation and describes whether the difference between the group responses on an item varies over that latent variable continuum (non-uniform DIF). Each of these steps provides a Chi-squared test statistic, which is used in the statistical test of DIF.

Just as a Chi-squared statistic is computed for each step in Zumbo's OLR DIF method, the corresponding effect size estimator is computed for each step. This corresponding effect size value is calculated as an R-squared from the OLR that can be applied to both binary and ordinal items. As Zumbo notes, because the R-squared is computed for the OLR, it has properties akin to the R-squared measures we commonly find in ordinary least-squares regression. Using these R-squared values, the magnitude of DIF can be computed by subtracting the R-squared value for the first step from that of the third step.

Finally, in order to classify an item as displaying DIF, one must consider both the Chi-squared test of DIF and the corresponding effect size measure. First, the two degrees of freedom Chi-squared test for DIF (i.e., testing for the gender and interaction effects simultaneously) must have a p-value less than or equal to 0.01. Second, the corresponding effect size measure must have a R-squared value of at least 0.035 – that is, Jodoin and Gierl's (2001) effect size criteria will be used to quantify the magnitude of DIF: R^2 values below 0.035 for negligible DIF, between 0.035 and 0.070 for moderate DIF, and above 0.070 for large DIF.

Furthermore, if DIF exists for an item, the steps computed in the calculation of DIF using Zumbo's (1999) ordinal logistic regression will be reviewed to determine if the DIF is uniform or non-uniform. Uniform DIF is determined by comparing the R-squared values between the second and first steps to measure the unique variation attributable to the group differences over-and-above the conditioning variable. If uniform DIF is found, the odds ratio will be used to interpret the direction of the DIF (i.e., are females or males more likely to endorse the item?). More precisely, the odds ratio can be used to determine the odds of one group responding higher to an individual item than those in the corresponding group, after matching on overall depressive symptomatology. The process is repeated for non-uniform DIF.

The reader should note that the DIF test is by definition a two-degree of freedom

test that involves an interaction and main effect. If one finds that this two-degree of freedom test is statistically significant than she/he may want to follow-up with the single-degree of freedom test to see whether the effect is predominantly uniform DIF.

Results from Ordinal Logistic Regression

We will provide the detailed computations for one of the items as a model for how to conduct DIF analyses. After conducting ordinal logistic regression on item 17: *I had crying spells*, the following model-fitting Chi-squares and Nagelkerke R-squares were found:

Step #1: Chi square = 190.63, df=1, R-squared=.338
Step #2: Chi square = 251.20, df=2, R-squared=.429
Step #3: Chi square = 254.77, df=3, R-squared=.434

This resulted in a two-degree freedom Chi-Square test (Step 3 – Step 1) of 64.14 with a p-value[1] of 0.0001 and a Nagelkerke R-squared (Step 3 – Step 1) of .096. Consequently, item 17 shows DIF with a large effect size.

Item 20: *I could not get going,* shows no DIF and a trivial to small effect size: model-fitting Chi-square = 0.25, df=2, p = 882, and Nagelkerke R-Squared = 0.0001.

Graphical Representation of DIF

The approach we will present is a non-parametric regression graphical visualization tool for DIF. The graphical representation of DIF is based on examining the relationship between total score and item responses for each group separately but simultaneously on the same graph.

A graphical display of items 20 and 17 are depicted in Figures 2 and 3. Note that (a) with reference to item response theory, each of these lines is, in essence, an item response function, and (b) we are looking for difference between the lines. As Zumbo and Hubley (2003) note, unlike the regression modeling approach, here we are not "covarying" out the total score. This graphical IRT approach is focusing on the area between the curves (in this case, the LOWESS curves) of the two groups. This is conceptually different than the logistic regression approach in that we are utilizing area between curves in determining DIF. This method does not technically match on the conditioning variable (total score).

As can be seen by Figure 2, there are no signs of DIF, indicating that item 20 is performing the same for males and females. As for item 17 in Figure 3, there are signs of DIF, which can be seen by the large area between the curves. This is indicating that item 17 is performing differently for males and females. In fact, males are not endorsing this item until one has a total scale score of 30.

Insert Figure 2 and 3 about here

---

[1] The p value is looked up from a probability table at the following website: <http://www.ubmail.ubalt.edu/~harsham/Business-stat/otherapplets/pvalues.htm#rchdist>

See Appendix B for a demonstration of how to use the drop-down windows in SPSS 11.5 to generate LOWESS smoothed regression lines like the ones seen in Figures 2 and 3.

Conclusion

The purpose of DIF is to rule out item bias. Item 20 showed no DIF, so there is no item bias. Item 17, however, did show DIF. Therefore; this item is displaying either bias or impact. By examining the R-squared results in Steps 1 through 3, one can see that the R-squared increases markedly from Steps 1 to 2, but far less so for Steps 2 to 3 in the ordinal logistic section above. This is interpreted to mean that the observed DIF is mostly uniform DIF. The presence of uniform DIF is supported by examining Figure 3, wherein the item response lines appear parallel for total scores greater than 30 – at less than 30, males are not endorsing the item so parallelism is not a question. This item is an interesting case of DIF because of the low endorsement at the low end of the scale.

In the end, it should be noted that one of the limitations of DIF analyses is that we are investigating differences between non-randomized groups. Like all such analyses, we therefore need to be cautious of interpretations of what is the source or "cause" of DIF. At this point all we know is that we have flagged item 17 as DIF, and Figure 3 shows us that males and females with the same level of depressive symptomatology (total score) may have different thresholds and hence endorse the item differently. In other words, is it that the difference lies in the reporting of depression (men and women report depression differently – item bias) or is it that women are truly more depressed than men (item impact)?

The graphical nonparametric approach to DIF is an important complement to the statistical modeling because, like all exploratory and data based nonparametric methods, it provides the analyst with a more robust view of the statistical problem and helps validate some of the assumptions -- such as, in our case, the monotonic increasing relationship between the total score and the item response. The nonparametric regression strategy (in this case, LOWESS smoothing) is robust in the same sense that the median is a more robust estimate of the center of a distribution as compared to the mean. Overall, the two methods together provide complementary evidence as to the presence of DIF.

In closing, two methodological notes are important. First, instead of matching on total score, one could match on factor scores (i.e., latent variable score). The advantage of this is that measurement error is less influential with a latent variable. Furthermore, for short scales, one may want to match on a 'rest score', which is the total score excluding the item under investigation. Each item would then, of course, have a different total score. This is called item-matching purification in the measurement literature. Finally, it is important to note that DIF analyses are conceptually different than scale-level analyses involving multi-group factor analysis. As Zumbo (2003) showed, scale level analyses ignore item-level DIF.

References

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 3*, 385-401.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. [Online]. Available: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html

Zumbo, B. D. (2003). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for Translating Language Tests. *Language Testing, 20,* 136-147.

Zumbo, B. D., & Hubley, A. M. (2003). Item Bias. In Rocío Fernández-Ballesteros (Ed.). *Encyclopedia of Psychological Assessment*, pp. 505-509. Sage Press, Thousand Oaks, CA.

Figure 1.
Instructions and Example Items from the CES-D

The CES-D

For each statement, circle the number (see the guide below) to indicate how often you felt or behaved this way **during the past week**.

0 = rarely or none of the time (less than 1 day)
1 = some or a little of the time (1-2 days)
2 = occasionally or a moderate amount of time (3-4 days)
3 = most or all of the time (5-7 days)

	Less than 1 day	1-2 days	3-4 days	5-7 days
2. I did not feel like eating; my appetite was poor.	0	1	2	3
17. I had crying spells.	0	1	2	3

Figure 2.
Graphic Model of no DIF: Item 20

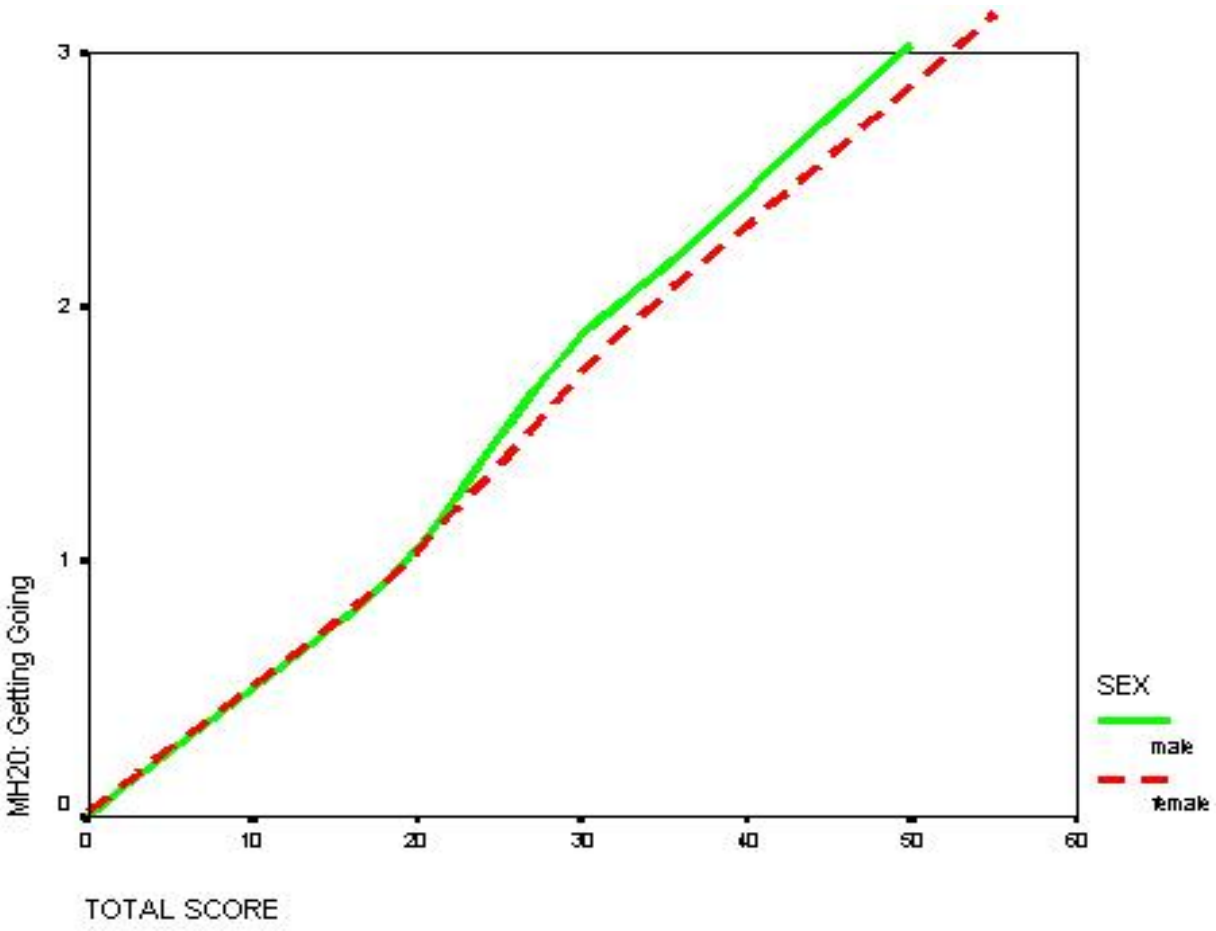
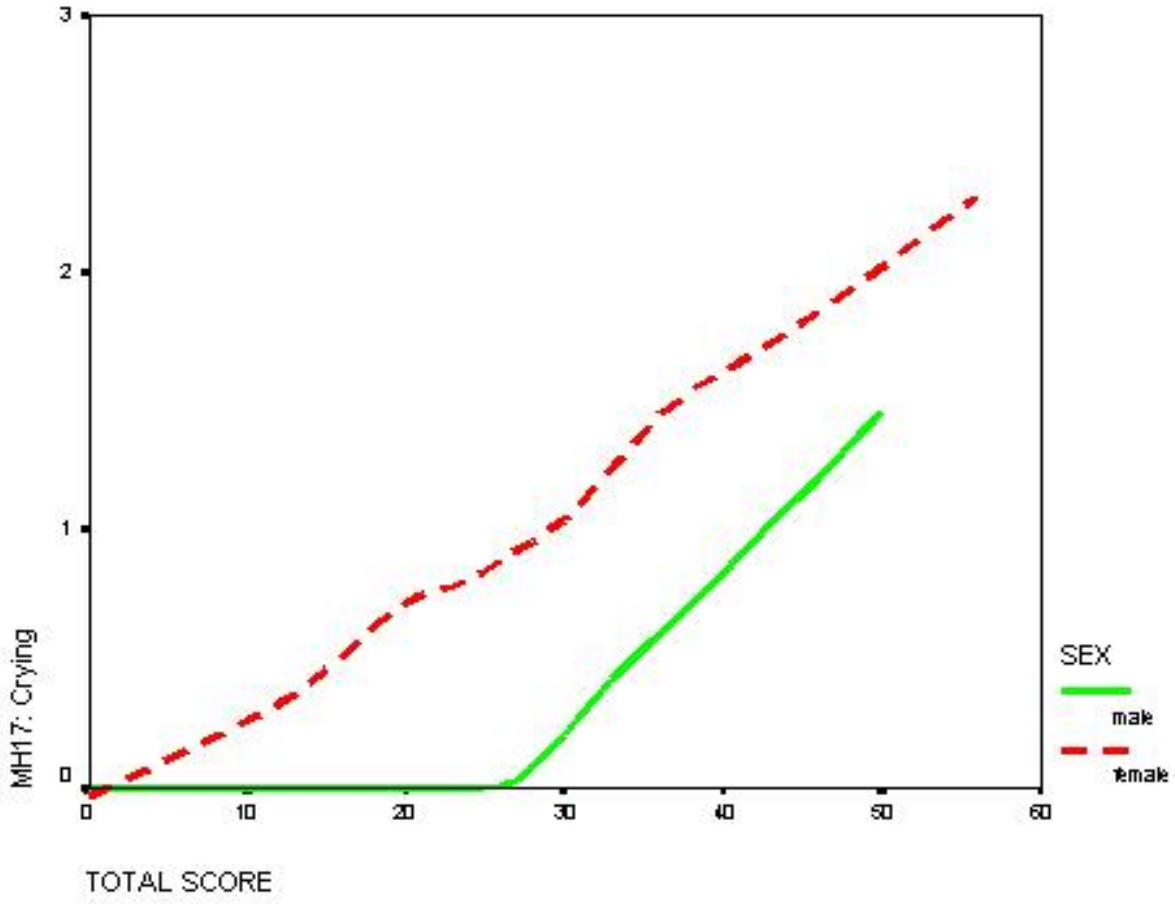


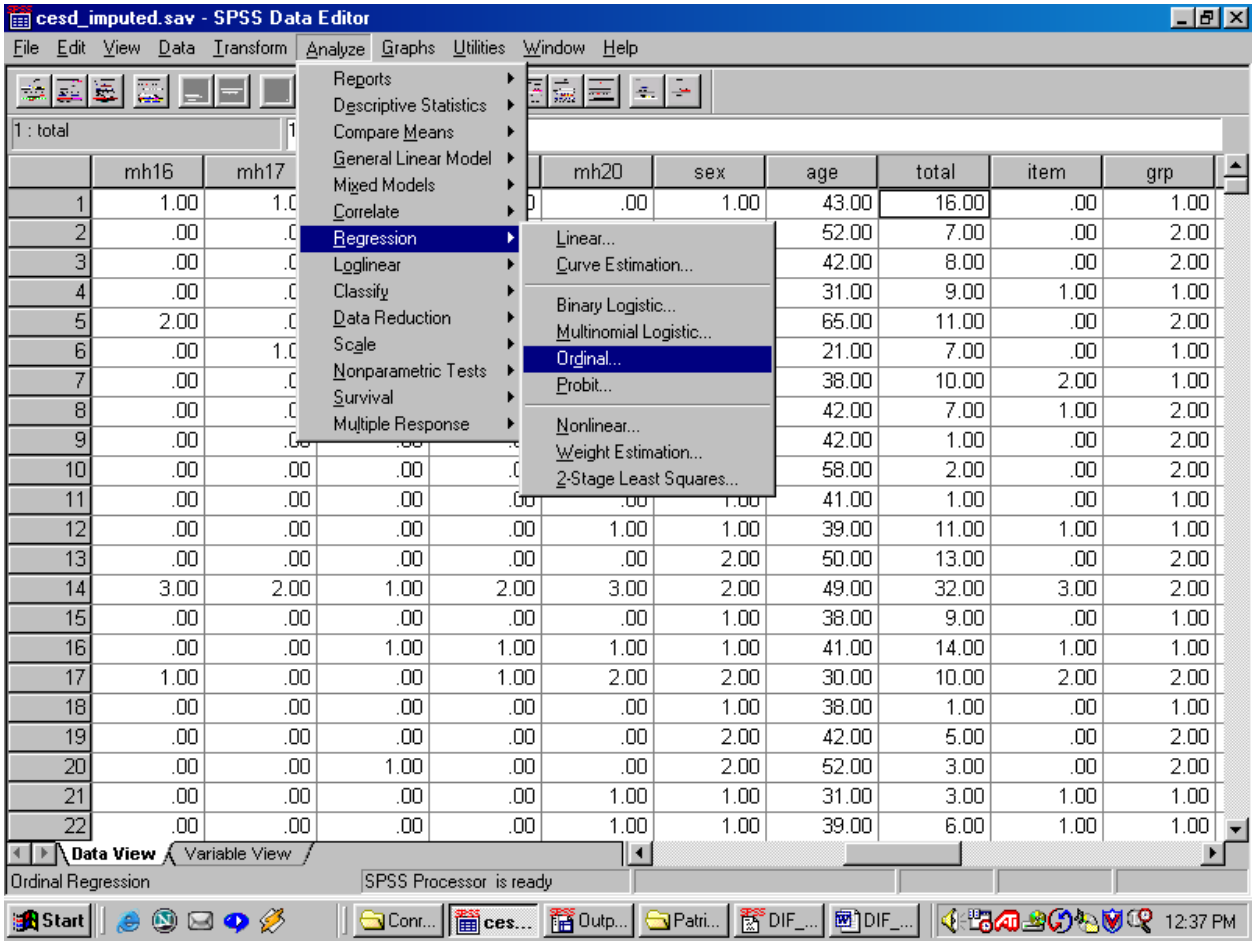
Figure 3.
Graphic Model of no DIF: Item 17



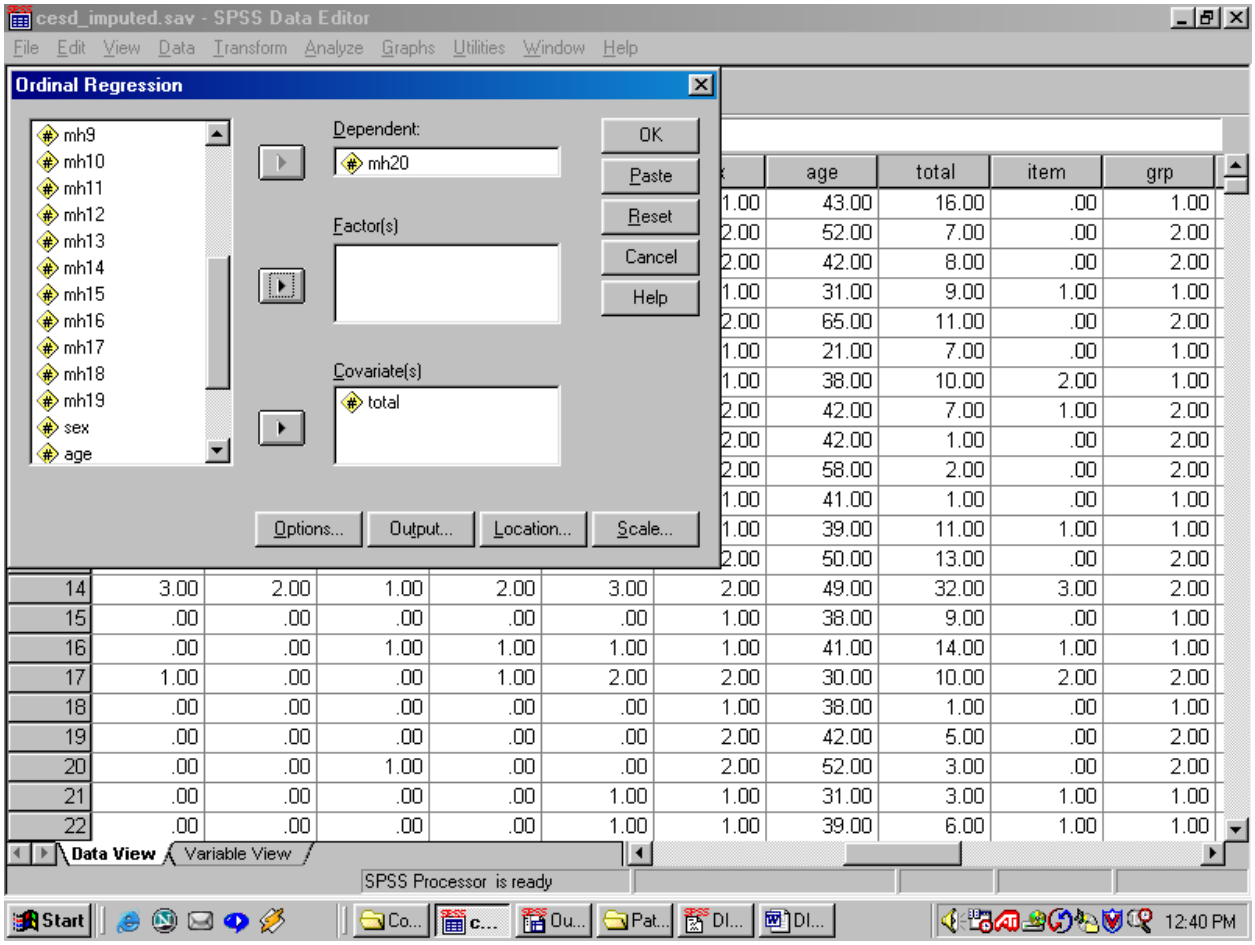
Appendix A

SPSS 11.5 Drop-down windows and syntax for Ordinal Logistic Regression DIF

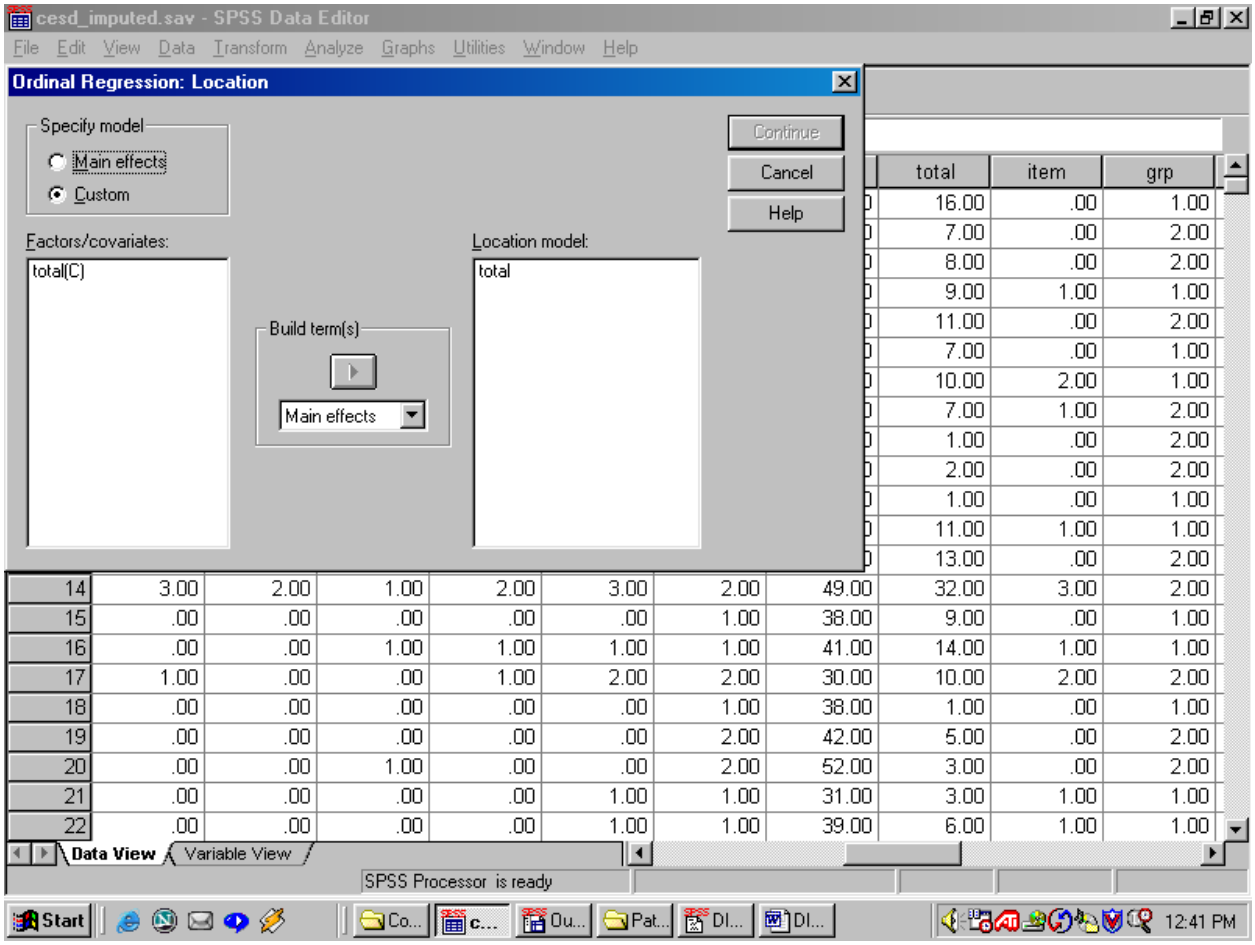
For all of the steps one begins here:



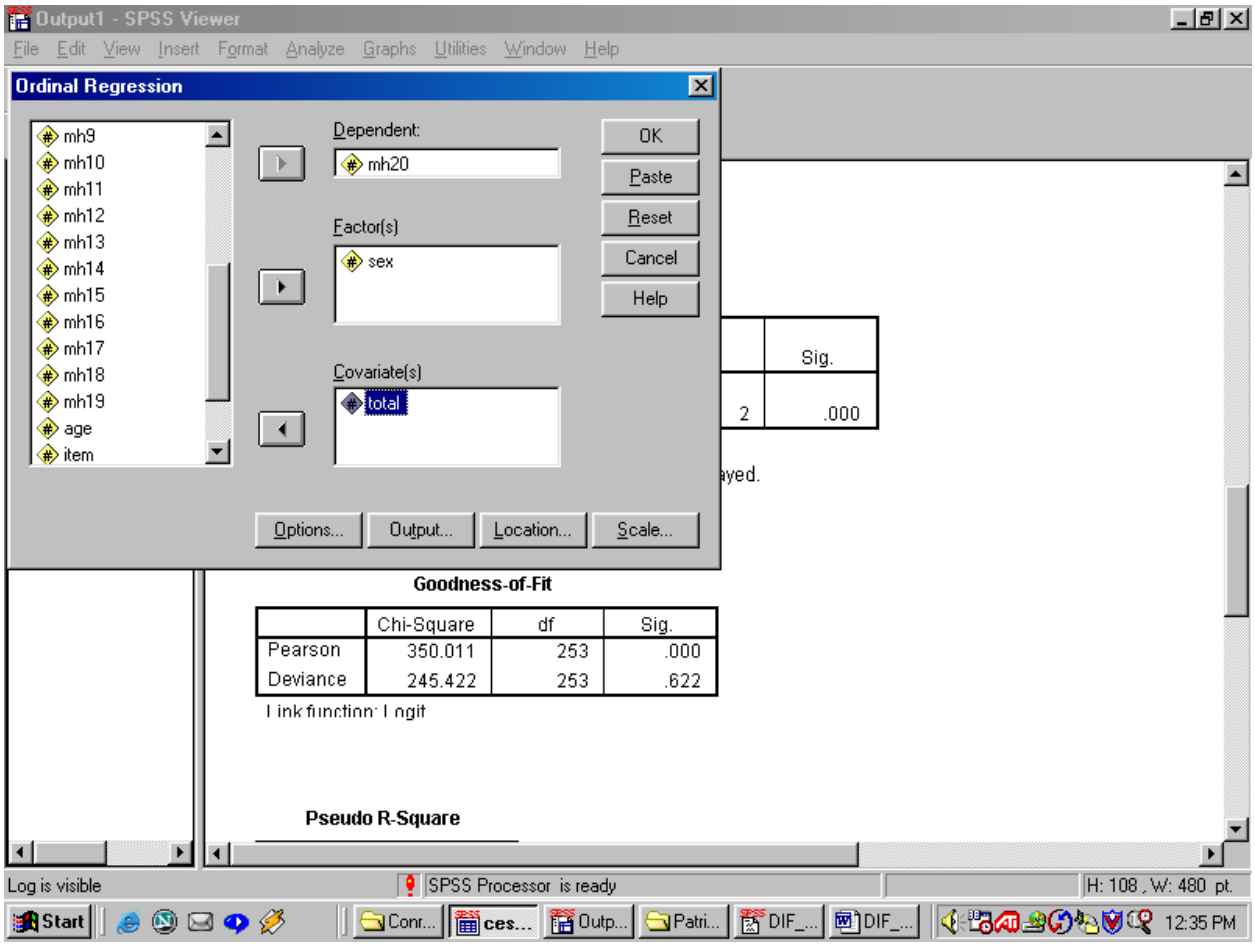
Step #1, the regression window looks like the following:



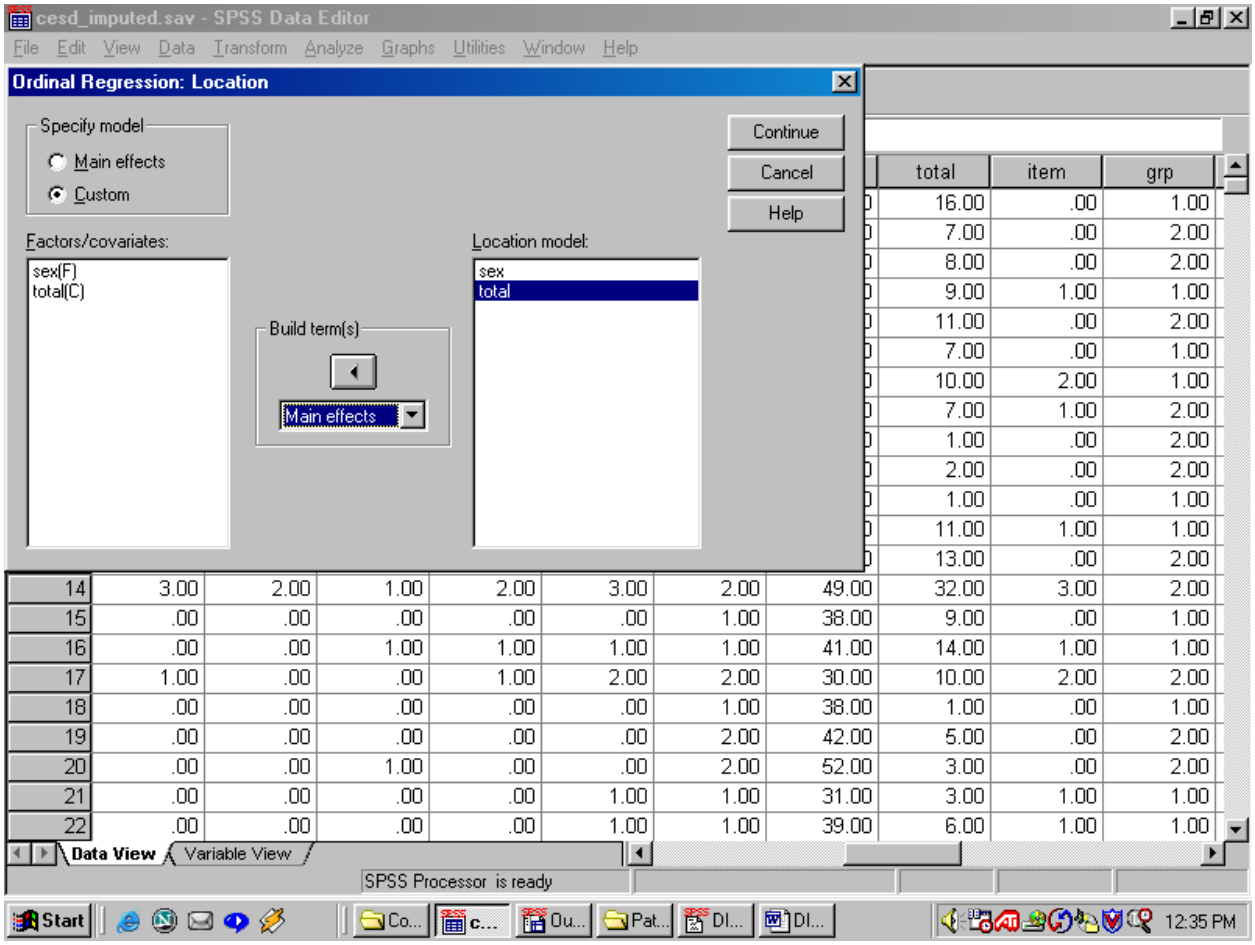
For Step #1, the location window looks like the following:



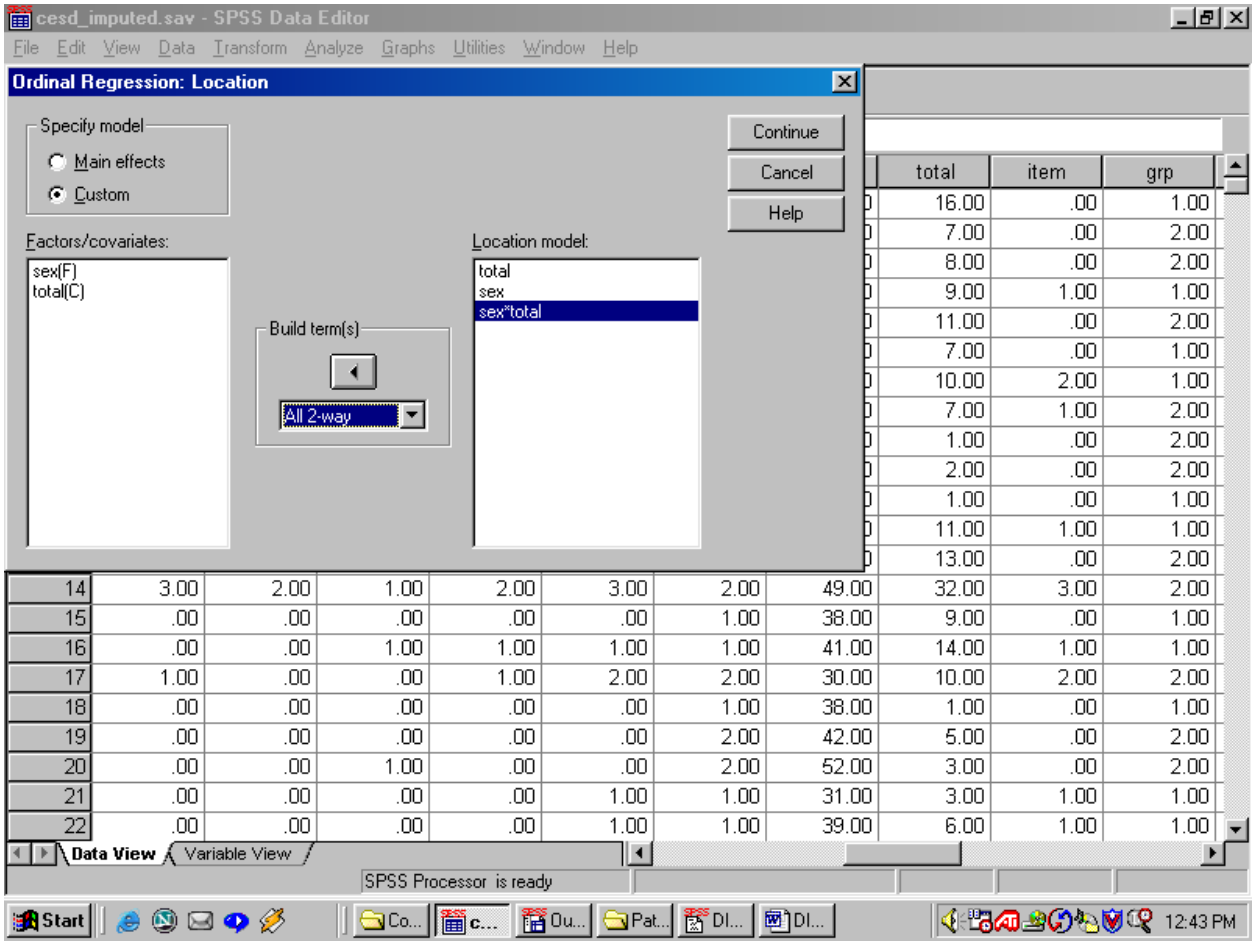
For Steps #2 and #3, the regression window looks like the following:



For Step #2, the Location window is:



For Step #3 the location window looks like the following:

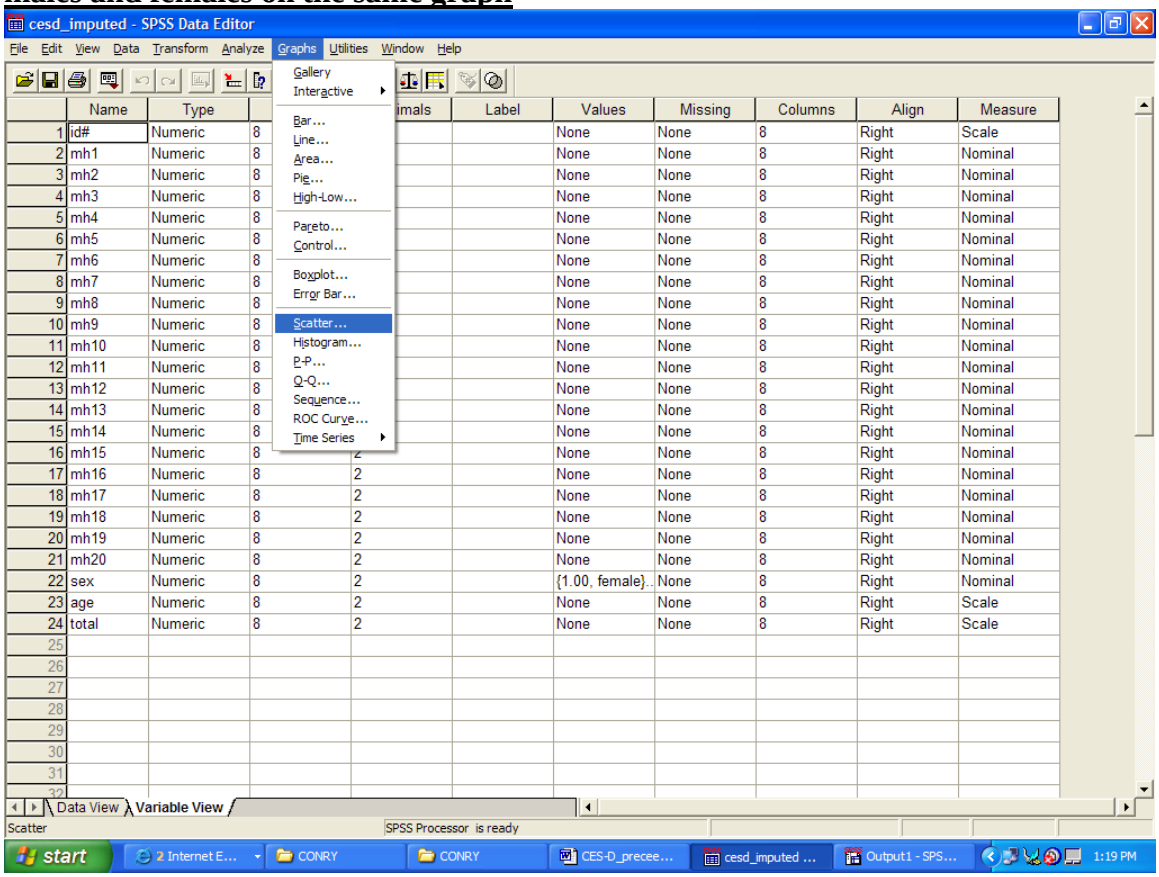


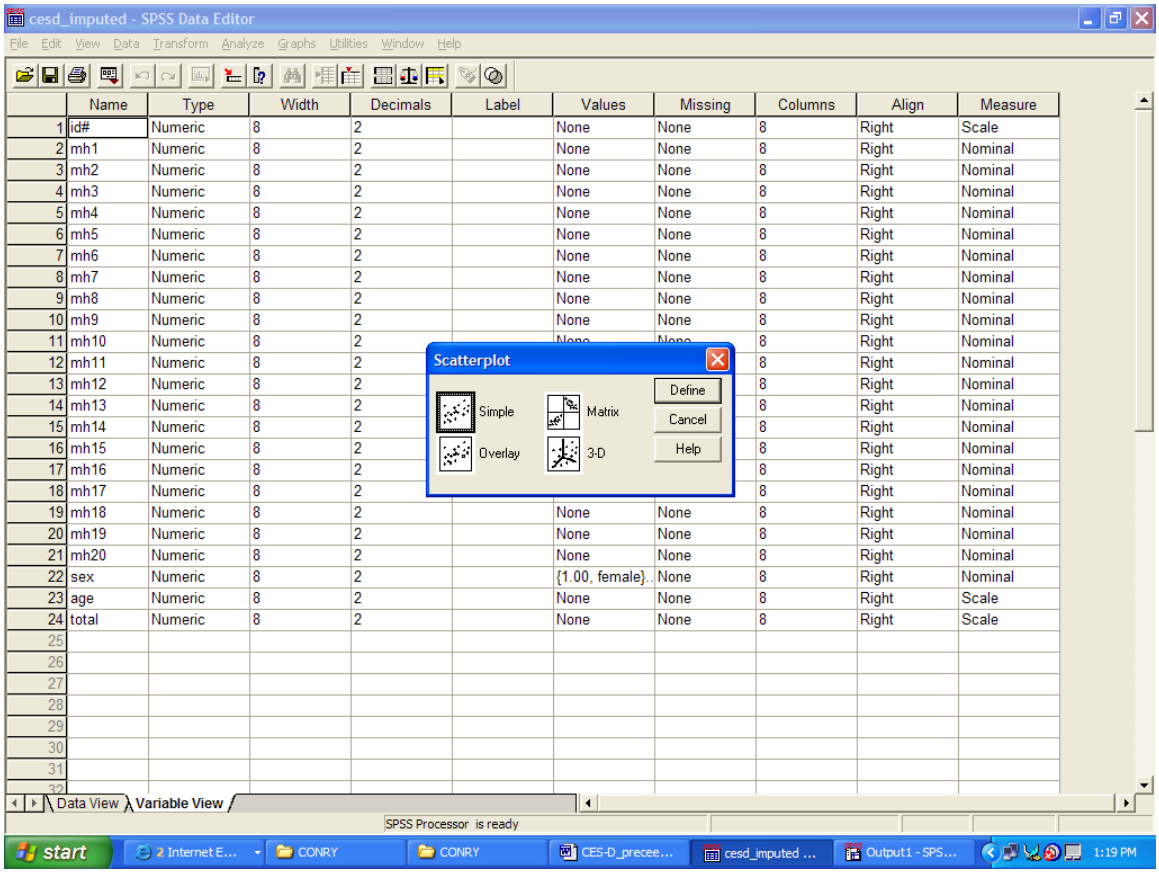
To conduct the analysis via syntax instead, use the following (the authors can send you this syntax as an attached file via e-mail):

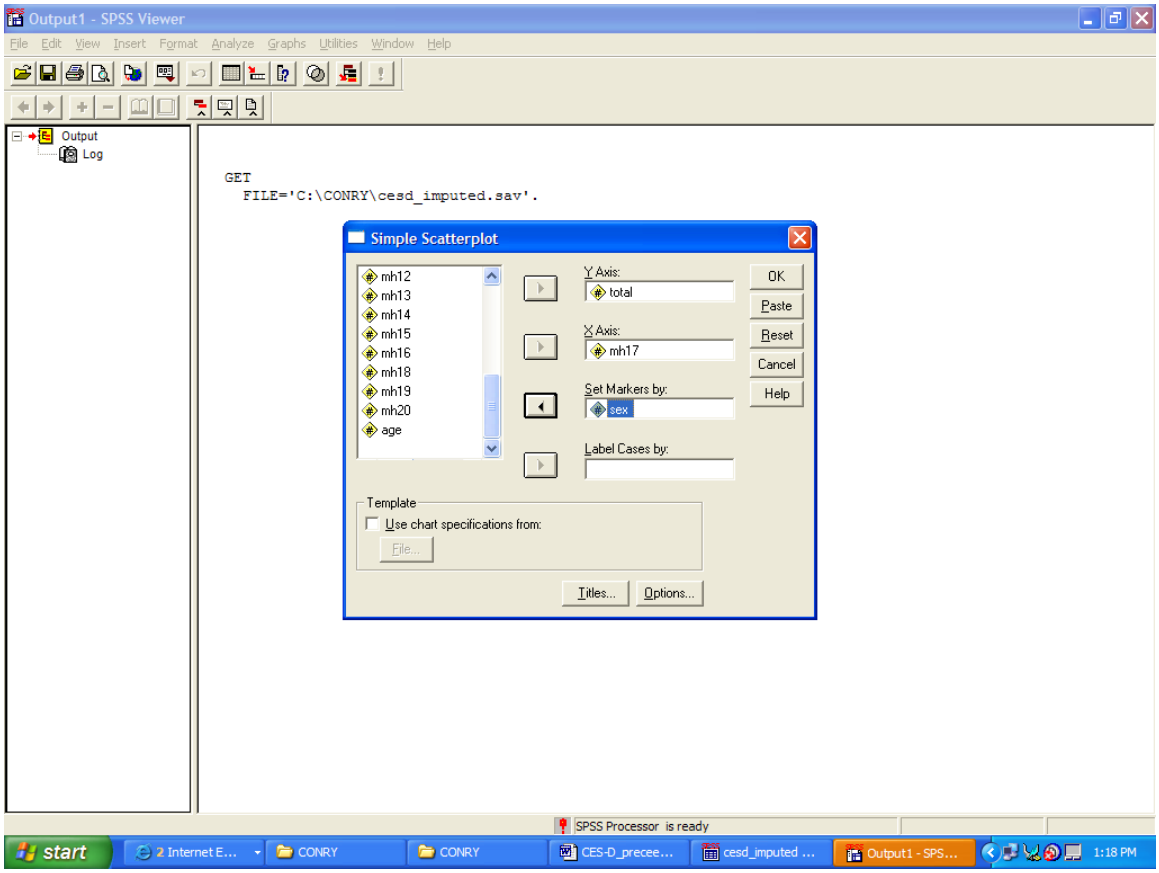
```
*        Progam:  DIF-SPSS,  2003.
*         Author:     Professor Bruno D. Zumbo, UBC.

* Step #1 Model with Total Score only.

* Only the three (3) compute commands need to be changed for you to
* use this syntax.

*  Insert the name of the item after "compute item=  ".
*  Insert the name of the total scale score after "compute total=  ".
*  Insert the name of the grouping variable after "compute grp=  ".

compute item=mh17.
compute total=total.
compute grp=sex.

PLUM
  item  WITH total
  /CRITERIA = CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
  PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
  /LINK = LOGIT
  /LOCATION =total
  /PRINT = HISTORY FIT PARAMETER SUMMARY KERNEL TPARALLEL.

* Step #2 Model with Total Score and Group effects.

PLUM
  item  BY grp  WITH total
  /CRITERIA = CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
  PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
  /LINK = LOGIT
  /LOCATION =grp total
  /PRINT = HISTORY FIT PARAMETER SUMMARY KERNEL TPARALLEL.

* Step #3 Model with Total Score, Group, and Interaction effects.


PLUM
  item  BY grp  WITH total
  /CRITERIA = CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
  PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
  /LINK = LOGIT
  /LOCATION =grp total grp*total
  /PRINT = HISTORY FIT PARAMETER SUMMARY KERNEL TPARALLEL.
```

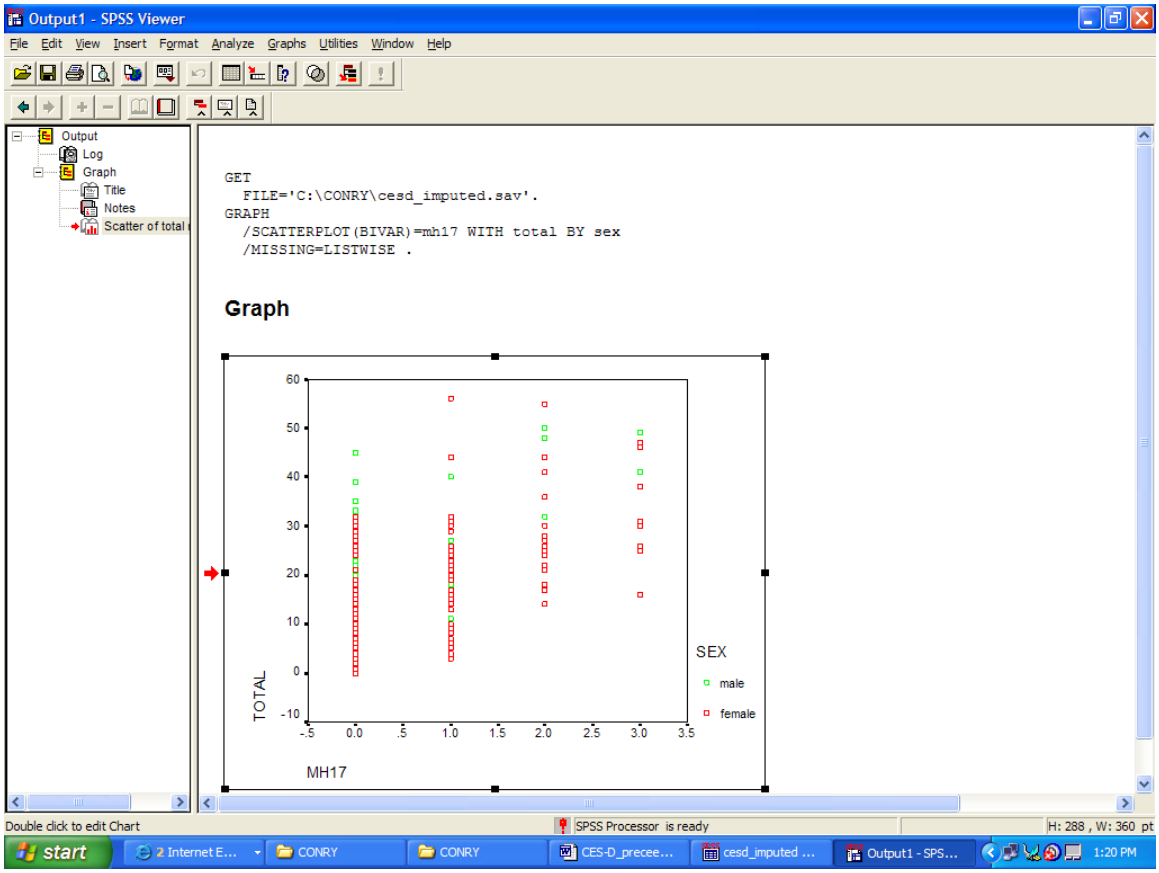
Appendix B

Drop-down windows for how to generate LOWESS smoothed regression lines for males and females on the same graph




cesd_imputed - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Name Type Width Decimals Label Values Missing Columns Align Measure
1 id# Numeric 8 2 None None 8 Right Scale
2 mh1 Numeric 8 2 None None 8 Right Nominal
3 mh2 Numeric 8 2 None None 8 Right Nominal
4 mh3 Numeric 8 2 None None 8 Right Nominal
5 mh4 Numeric 8 2 None None 8 Right Nominal
6 mh5 Numeric 8 2 None None 8 Right Nominal
7 mh6 Numeric 8 2 None None 8 Right Nominal
8 mh7 Numeric 8 2 None None 8 Right Nominal
9 mh8 Numeric 8 2 None None 8 Right Nominal
10 mh9 Numeric 8 2 None None 8 Right Nominal
11 mh10 Numeric 8 2 8 Right Nominal
12 mh11 Numeric 8 2 8 Right Nominal
13 mh12 Numeric 8 2 8 Right Nominal
14 mh13 Numeric 8 2 8 Right Nominal
15 mh14 Numeric 8 2 8 Right Nominal
16 mh15 Numeric 8 2 8 Right Nominal
17 mh16 Numeric 8 2 8 Right Nominal
18 mh17 Numeric 8 2 8 Right Nominal
19 mh18 Numeric 8 2 None None 8 Right Nominal
20 mh19 Numeric 8 2 None None 8 Right Nominal
21 mh20 Numeric 8 2 None None 8 Right Nominal
22 sex Numeric 8 2 {1.00, female}.. None 8 Right Nominal
23 age Numeric 8 2 None None 8 Right Scale
24 total Numeric 8 2 None None 8 Right Scale
Scatterplot
Simple
Matrix
Overlay
3-D
Define
Cancel
Help
Data View Variable View
SPSS Processor is ready
start
2 Internet E...
CONRY
CONRY
CES-D_precee...
cesd_imputed ...
Output1 - SPS...
1:19 PM


Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Output
Log
GET
FILE='C:\CONRY\cesd_imputed.sav'.
Simple Scatterplot
mh12
mh13
mh14
mh15
mh16
mh18
mh19
mh20
age
Y Axis:
total
X Axis:
mh17
Set Markers by:
sex
Label Cases by:
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Titles...
Options...
SPSS Processor is ready
start
2 Internet E...
CONRY
CONRY
CES-D_precee...
cesd_imputed ...
Output1 - SPS...
1:18 PM



Double click on the graph to bring up Chart Editor and request LOWESS smoothed regression lines for males and females.

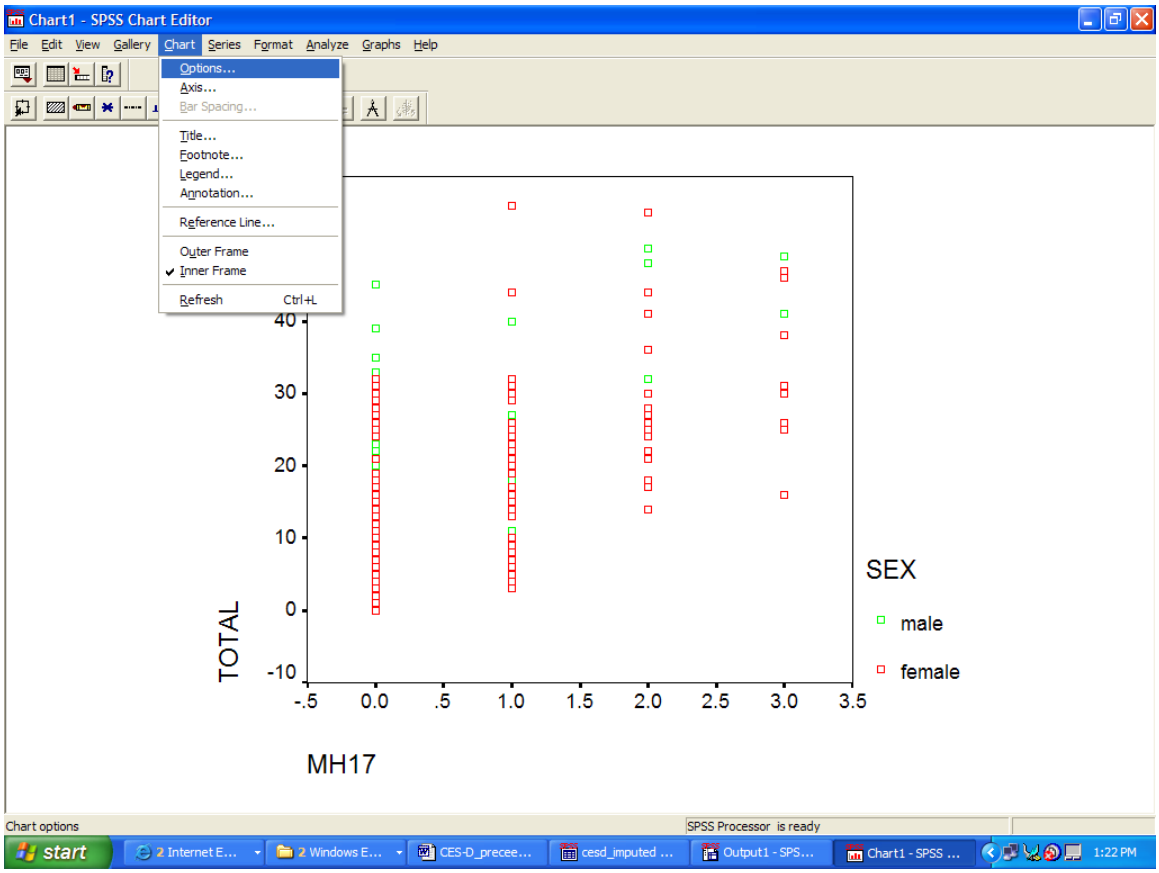


Chart1 - SPSS Chart Editor
File Edit View Gallery Chart Series Format Analyze Graphs Help
Options...
Axis...
Bar Spacing...
Title...
Footnote...
Legend...
Annotation...
Reference Line...
Outer Frame
Inner Frame
Refresh Ctrl+L
TOTAL
MH17
SEX
male
female
Chart options
SPSS Processor is ready
start
1:22 PM

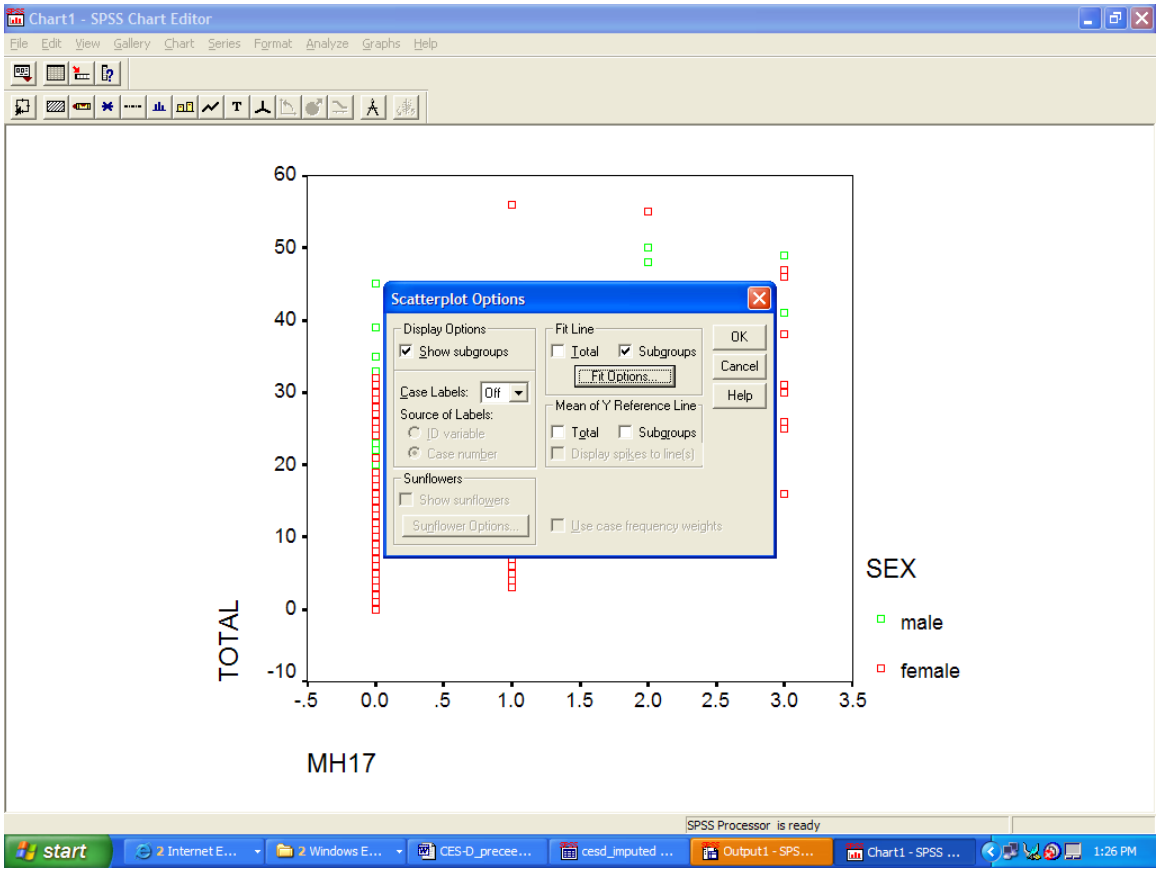


Chart1 - SPSS Chart Editor
File Edit View Gallery Chart Series Format Analyze Graphs Help
Scatterplot Options
Display Options
Show subgroups
Case Labels: Off
Source of Labels:
ID variable
Case number
Sunflowers
Show sunflowers
Sunflower Options...
Fit Line
Total
Subgroups
Fit Options...
Mean of Y Reference Line
Total
Subgroups
Display spikes to line(s)
Use case frequency weights
OK
Cancel
Help
TOTAL
MH17
SEX
male
female
SPSS Processor is ready
start
1:26 PM

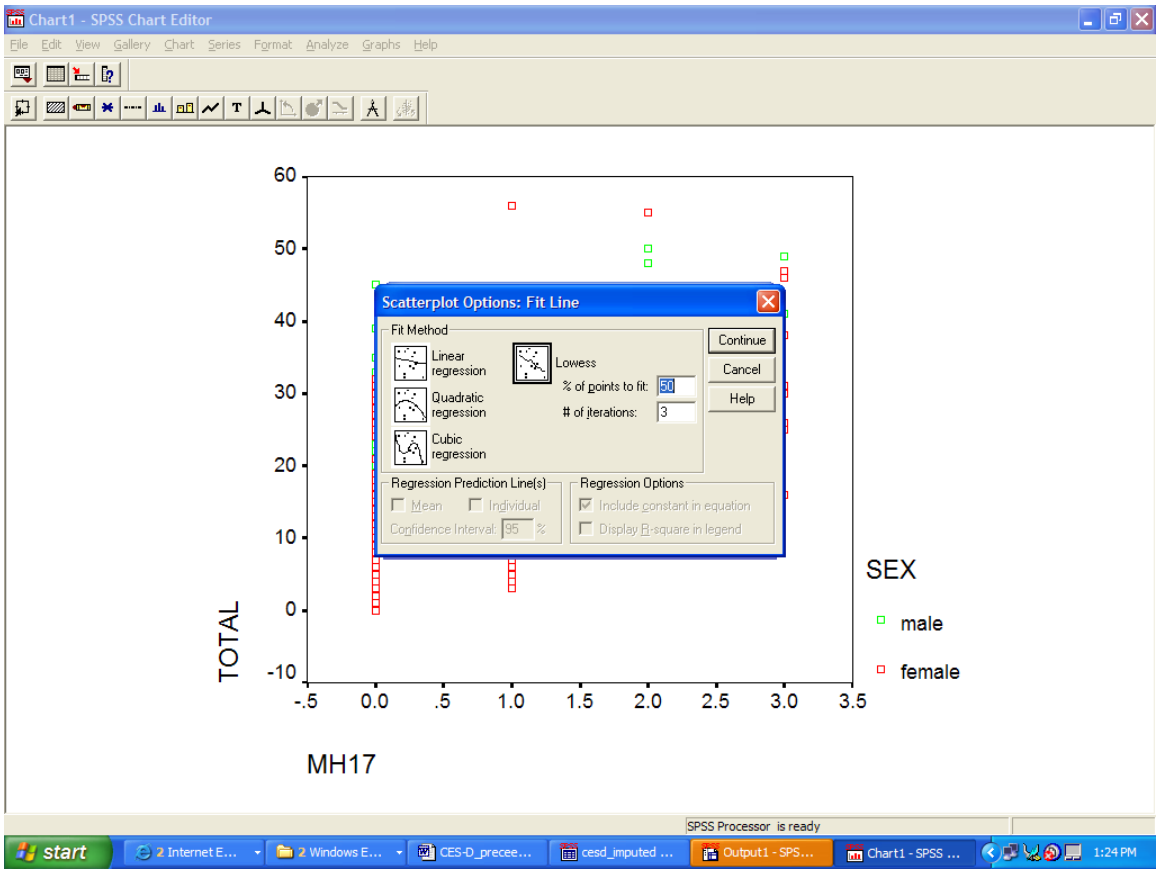


Chart1 - SPSS Chart Editor
File Edit View Gallery Chart Series Format Analyze Graphs Help
Scatterplot Options: Fit Line
Fit Method
Linear regression
Quadratic regression
Cubic regression
Lowess
% of points to fit: 50
# of iterations: 3
Continue
Cancel
Help
Regression Prediction Line(s)
Mean
Individual
Confidence Interval: 95 %
Regression Options
Include constant in equation
Display R-square in legend
TOTAL
MH17
SEX
male
female
SPSS Processor is ready
start
1:24 PM