CCRAM Technical Report 022-03

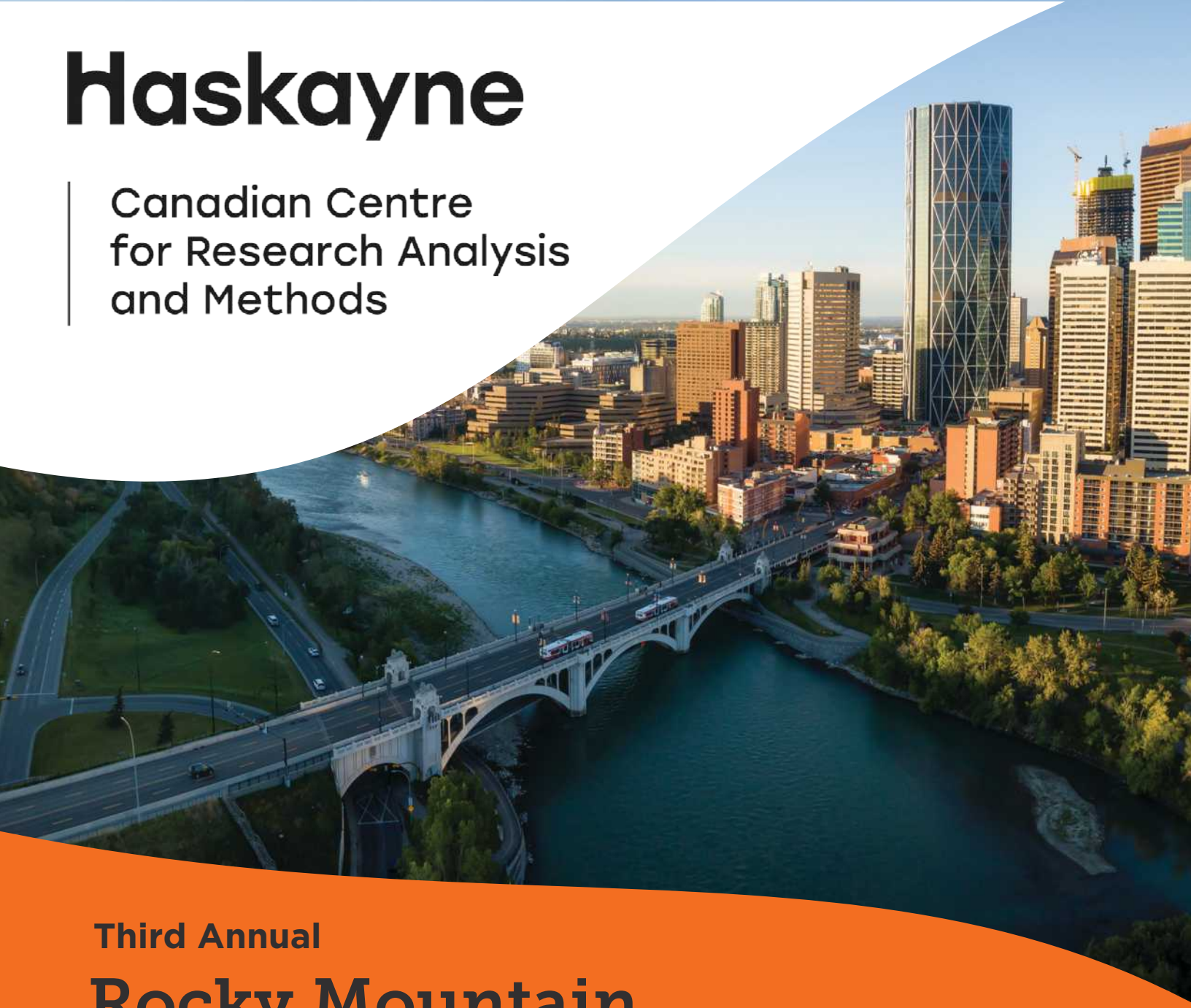
Changing the Reference Group when Using Indicator Coding for a Multicategorical Variable in PROCESS

Andrew F. Hayes
Haskayne School of Business
University of Calgary

haskayne School of Business | Canadian Centre for Research Analysis and Methods

Haskayne

Canadian Centre for Research Analysis and Methods



Third Annual

Rocky Mountain Methodology Academy

July 15 – July 22, 2025

The Canadian Centre for Research Analysis and Methods hosts its annual Rocky Mountain Methodology Academy in Calgary, Alberta, Canada. Choose from ten courses taught by experts in social science data analysis and research methods. Between sessions and after class, explore Calgary and attend additional free lectures and events, including a day trip to Banff in the Canadian Rockies.

UNIVERSITY OF CALGARY

Session 1: July 15-16, 2025

Longitudinal Data Analysis and Visualization
DR. ANDREA HOWARD, PHD (CARLETON UNIVERSITY)

Data are often collected longitudinally, meaning the same variables are measured repeatedly over time, with the goal of understanding how variables change within and between people over time. This course provides a broad overview of various methods of quantifying, modeling and visualizing change in variables over time and how to test hypotheses about intraindividual and interindividual change.

Introduction to Social Network Analysis
DR. JENNY GODLEY, PHD (UNIVERSITY OF CALGARY)

Social network analysis examines the patterning of relationships between individuals and groups to understand social action. This course will cover the design, collection, analysis and interpretation of both whole and ego-centred network data.

Mediation Analysis
DR. ANDREW F. HAYES, PHD (UNIVERSITY OF CALGARY)

Mediation analysis is among the most widely used data analysis techniques in the social, health, and business sciences. In this course, you will learn about the fundamentals of estimation and inference about direct, indirect, and total effects and how mediation analysis is used to study the mechanism(s) by which effects operate. Application focuses on use of the popular PROCESS macro for SPSS, SAS, and R invented by the course instructor.

Intersession 1: July 17, 2025

There is no additional charge to attend these events. A reception will follow the keynote address.

Guest Lecture
DR. AMANDA MONTOYA, UNIVERSITY OF CALIFORNIA LOS ANGELES

Keynote Address
DR. ANDREA HOWARD, CARLETON UNIVERSITY

Session 2: July 18-19, 2025

Introduction to Structural Equation Modeling
DR. DOUG BAER, PHD (UNIVERSITY OF VICTORIA)

This course introduces the fundamentals of structural equation modeling as a general analytical tool, including how to set up measurement and structural models, latent variables, path analysis, definitions and quantification of model fit and the implementation of structural equation modeling in statistical software.

Interactions in Regression Analysis
DR. ANDREW F. HAYES, PHD (UNIVERSITY OF CALGARY)

Effects that scientists study rarely are uniform across people, context, or stimuli. Moderation analysis is used to examine the extent to which an effect depends on another variable, meaning it is moderated or the two variables interact. In the context of regression analysis and emphasizing the PROCESS macro invented by the instructor, this class covers the fundamentals of moderation analysis, including model estimation, interpretation, visualization, and probing interactions.

Experience Sampling and Implementation
DR. SABRINA THAI, PHD (BROCK UNIVERSITY)

Experience sampling methods are a powerful approach that allows researchers to examine how psychological phenomena unfold in daily life. This course introduces experience sampling methods and provides an overview of issues to be considering in using these designs, including signal frequency, sample size, power, question wording, compensation and recruitment, data handling, and technology implementation.

Intersession 2: July 20, 2025

Canadian Rockies Day Trip

Included with your registration while space is available, enjoy a day trip to Banff, Alberta, one of the most popular tourist destinations in Canada. Seats are limited, so register for the academy now to reserve your space on the bus.

Session 3: July 21-22, 2025

Introduction to Multilevel Modeling
DR. JASON RIGHTS, PHD (UNIVERSITY OF BRITISH COLUMBIA)

This course provides an introduction to multilevel modeling, with a focus on its application within the social, education, health and business sciences. Participants will learn fundamental statistical principles underlying multilevel modeling, a variety of techniques and methods that can be used in many different research contexts and how to appropriately specify models and interpret results in practice.

Scale Development and Psychometrics
DR. JESSICA FLAKE, PHD (UNIVERSITY OF BRITISH COLUMBIA)

Researchers in the academic and private sectors often need to measure attitudes, intentions, satisfaction, or motivation. Because scale scores are used to make decisions and evaluate research outcomes, develop a product, or hire or promote an employee, researchers need to thoroughly evaluate their validity. This course covers how to develop, evaluate, and refine scales using modern psychometric methods.

Introduction to Mixed Methods Research
DR. CHERYL POTH, PHD (UNIVERSITY OF ALBERTA)

Mixed methods research requires specific integration of knowledge and skills that also leverage existing qualitative and quantitative skills. Participants in this course will learn how to distinguish credible mixed methods research and have opportunities to ask questions about recent integration practice advancements. Discussions of the many perceived (and real) integration challenges when designing, executing and disseminating mixed methods research will provide foundational understandings for participants to engage in the design of their own mixed methods research projects.

Latent Profile Analysis
DR. MATTHEW MCLARNON, PHD (MOUNT ROYAL UNIVERSITY)

Latent profile analysis is a family of statistical models that can be used to identify unobserved, heterogenous and qualitatively distinct subgroups in one's data. This course will provide participants with the theoretical and conceptual background and applied analytical skills needed to specify an appropriate analytical model, interpret the results and thoroughly address research questions using latent profile analysis.

Register now! Seats are limited.

To register for courses and for more information, visit **haskayne.ucalgary.ca/CCRAM/academy**

All courses will take place on the main campus of the University of Calgary.

While space is available, your registration includes a day trip by chartered bus to the town of Banff in the Canadian Rockies.

The more courses you attend the more you save!

- → Courses are $950 (CAD) each
- → Enroll in 2 courses: $1,800 (CAD) **SAVE 5%**
- → Enroll in 3 courses: $2,500 (CAD) **SAVE 10%**
- → Longitudinal Toolkit: $2,280 (CAD) **SAVE 20%**
- → Graduate students are eligible for an **additional 10% discount.**

Note: As courses run from 9am to 5pm each day, you can register for only one course per session. The Longitudinal Toolkit includes longitudinal data analysis and visualization, experience sampling and implementation, and introduction to multilevel modeling. Prices above do not include 5% federal goods and services tax (GST). Consult your bank for current exchange rates.

HASKAYNE SCHOOL OF BUSINESS
2500 University Drive NW
Calgary, AB T2N 1N4, CANADA

Canadian Centre for Research Analysis and Methods (CCRAM)

haskayne.ucalgary.ca/CCRAM
ccram@ucalgary.ca

Changing the Reference Group When Using Indicator Coding for a Multicategorical Variable in PROCESS

Andrew F. Hayes
University of Calgary Haskayne School of Business

A multicategorical variable (a variable representing membership in one of three or more groups) is typically a single variable in a data set containing a set of numerical codes representing which group a case in the data belongs in. Such a variable cannot be used as a predictor in a regression analysis as is. Doing so will usually produce nonsense. The proper approach to including a multicategorical variable representing *k* groups on the right-hand side of a regression equation is to use a set of $k - 1$ variables that code group membership. There are many coding systems that can be used. One of the most popular is *indicator* coding, also called *dummy* coding. With indicator coding, one group is chosen as the *reference group* and cases in that group receives a value of 0 on the $k - 1$ indicator or dummy variables representing the *k* groups. The cases in the remaining $k - 1$ groups are each given a value of 1 on the indicator variable for that group and all other indicator variables for cases in that group are set to 0.

For example, suppose in your data file you have a variable named `group`, with the $k = 4$ groups coded with values 1, 2, 3, and 4 in the data. The table below represents a possible indicator coding system:

Table 1.

group	X1	X2	X3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

In this system, X1, X2, and X3 are the indicator variables, and group 4 is the reference group because it is represented with zeros on all 3 of the indicator variables. Notice that although there are only three indicator variables, there are four patterns of zeros and ones, one pattern for each group. It is the pattern of values on X1, X2, and X3 that represents group membership.

Using the **`mcx`**, **`mcw`**, or **`mcz`** options in PROCESS, you can tell PROCESS to represent a multicategorical variable specified as X, W, or Z in the model with one of several preprogrammed coding systems. Indicator coding is available by specifying option 1, as discussed in *Introduction to Mediation, Moderation, and Conditional Process Analysis*. When using option 1, PROCESS will always use the group with the numerically *smallest* code as the reference group. For example, in the four-group example above, PROCESS would choose the group coded 1 as the reference group because "1" is the smallest value on the group variable.

To illustrate, we use the data from Chapter 6 of *Introduction to Mediation, Moderation, and Conditional Process Analysis*. In the data file, the three experimental conditions are coded 0, 1, and 2 in the variable named `protest` for the "no protest," "individual protest," and "collective protest" conditions, respectively. As the "no protest" group has the numerically smallest code (0) in the `protest` variable, PROCESS will treat it as the reference group when creating the indicator codes. The PROCESS command below (SPSS first, SAS second, R third) estimates evaluation of the attorney

from protest condition, specifying protest as a multicategorical variable and using indicator coding to represent the three groups. (Note that this code will only work with PROCESS version 4.1 or later, as model 0 was not available as a model number until version 4.1 released in April 2022).

```
process y=liking/x=protest/model=0/mcx=1.
```

```
%process(data=protest,y=liking,x=protest,model=0,mcx=1)
```

```
process(data=protest,y="liking",x="protest",model=0,mcx=1)
```

This produces as output

```
Model  : 0
    Y  : liking
    X  : protest

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest      X1      X2
    .000    .000    .000
   1.000   1.000    .000
   2.000    .000   1.000

**************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.3102      .1614    32.9083      .0000     4.9909     5.6296
X1            .5158      .2255     2.2870      .0239      .0695      .9621
X2            .4431      .2231     1.9863      .0492      .0016      .8845
```

Notice at the beginning of the output that PROCESS has produced a table representing the indicator coding system used. The no protest group is the reference group, as X1 and X2 are both 0 for this group. The *F*-ratio in the model summary is identical to the *F*-ratio from a single factor or "one-way" analysis of variance testing the null hypothesis that the group means are the same. Here, $F(2,126) = 3.055$, $p = .051$.

In the data, the group means are 5.310 for the no protest group, 5.826 for the individual protest group, and 5.753 for the collective protest group. The model is $\hat{Y} = 5.310 + 0.516X_1 + 0.443X_2$. When the pattern of indicator codes for each group are plugged into the regression model, the model produces an estimate of *Y* for each group that corresponds to that group's mean on *Y*:

No protest group: $\hat{Y} = 5.310 + 0.516(0) + 0.443(0) = 5.310$

Individual protest group: $\hat{Y} = 5.310 + 0.516(1) + 0.443(0) = 5.826$

Collective protest group: $\hat{Y} = 5.310 + 0.516(0) + 0.443(1) = 5.753$

The regression constant is the mean for the no protest reference group. The regression weight for X1 is the difference between means of the individual protest group and the no protest reference group: 0.516 = 5.826 – 5.310. And the regression weight for X2 is the difference between the means of the collective protest group and the no protest reference group: 0.443 = 5.753 – 5.310. Both the differences are statistically significant. So each of these regression coefficients quantify one group's mean relative to the no protest group mean. That's why the no protest group is called the reference group here.

Changing the Reference Group

But what if you want to a different group as the reference? As PROCESS is programmed to always make the group with the numerically smallest group the reference, you will need to do something to change the reference group. The most obvious solution is to recode the groups so that that the group you desire as the reference has the numerically smallest code. For example, if you want the collective protest group (`protest` = 2) to be the reference, you could swap the 0 and 2 codes for the no protest and collective protest groups. So you don't lose the original coding, I recommend putting the new codes in a different variable.

Suppose the new codes were held in a variable called `protest2` with protest2 = 0, 1, and 2 for the collective protest, individual protest, and no protest groups, respectively. Then the PROCESS command above, but using `protest2` for *X*, generates

```
Model  : 0
    Y  : liking
    X  : protest2

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest2       X1       X2
     .000     .000     .000
    1.000    1.000     .000
    2.000     .000    1.000

**************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.7533      .1540    37.3530      .0000     5.4485     6.0581
X1            .0727      .2203      .3300      .7419     -.3633      .5088
X2           -.4431      .2231    -1.9863      .0492     -.8845     -.0016
```

Now the regression coefficient for X1 quantifies the difference between the individual protest group mean and the collective protest group mean: 0.073 = 5.826 – 5.753, and the regression coefficient for X2 quantifies the difference between the no protest group mean and the collective protest group mean: –0.443 = 5.310 – 5.753. The regression constant is the mean for the collective protest group, which is our new reference group. The regression model still generates estimates of *Y* for each group that correspond to that group's mean *Y*:

Collective protest group: $\hat{Y} = 5.753 + 0.073(0) - 0.443(0) = 5.753$
Individual protest group: $\hat{Y} = 5.753 + 0.073(1) - 0.443(0) = 5.826$
No protest group: $\hat{Y} = 5.753 + 0.073(0) - 0.443(1) = 5.310$

The test of equality of the three groups means is not affecting by changing the reference group. It is still $F(2,126) = 3.055$, $p = .051$.

An alternative approach to changing the reference group is to program your own indicator coding system using the **xcatcode** option described in Appendix A of *Introduction to Mediation, Moderation, and Conditional Process Analysis*. The desired coding system with the collective protest group as the reference is

Table 2.

	X1	X2
No protest (`protest` = 0)	1	0
Individual protest (`protest` = 1)	0	1
Collective protest (`protest` = 2)	0	0

In Table 2, the groups in the rows are in the same order as the groups are represented with ascending numbers in the variable coding groups that is being used in the analysis. In this case, groups are stored in the `protest` variable with values 0, 1, and 2. So the first row is the group coded 0, the second row is the group coded 1, and the third row is the group coded 2. Getting the order right is important because PROCESS expects this order when it creates the variable codes in response to your use of the **xcatcode** option.

To tell PROCESS to use the coding system in Table 2, specify **mcx** option 5 in the PROCESS command and then list the numerical codes in the table as a sequence of numbers following the **xcatcode** option, reading the numbers in Table 2 from left to right, top to bottom as you enter the sequence. Thus, in SPSS, SAS, and R, respectively, the PROCESS command is

```
process y=liking/x=protest/mcx=5/model=0/xcatcode=1,0,0,1,0,0.
```

```
%process (data=protest,y=liking,x=protest,model=0,mcx=5,xcatcode=1 0 0 1 0 0)
```

```
process (data=protest,y="liking",x="protest",model=0,mcx=5,xcatcode=c(1,0,0,1,0,0))
```

This command generates the output below, which of course is identical to what was generated when we just recoded the groups instead of programming our own indicator codes. Notice that the table toward the top of the output shows that the group coded 2 in `protest` (the collective protest group) is now the reference group. It is important to check this table to make sure that you entered the sequence of numbers correctly when using the **`xcatcode`** option.

```
Model  : 0
    Y  : liking
    X  : protest

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest      X1      X2
    .000   1.000    .000
   1.000    .000   1.000
   2.000    .000    .000

**************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.7533      .1540    37.3530      .0000     5.4485     6.0581
X1           -.4431      .2231    -1.9863      .0492     -.8845     -.0016
X2            .0727      .2203      .3300      .7419     -.3633      .5088
```

Reversing the Differencing

The regression coefficients for the indicator variables quantify a difference in estimated values of *Y* between one of the groups and the reference group. As discussed earlier, these can be interpreted as a difference between group means, where the reference group mean is subtracted from the mean of the other groups. If the reference group mean is smaller, the regression coefficient is positive. If the reference group mean is larger, the coefficient is negative.

There may be occasions where you might prefer that the subtraction be reversed, such that one of the group means is subtracted from the reference group mean. This won't change the substantive interpretation of the difference between the group means, of course, but it will flip the signs of the corresponding regression coefficient and endpoints of a confidence interval. The standard error for the regression coefficient and the *p*-value for testing the null of no difference will be the same, however.

The reversal of the subtraction and corresponding reversal of the regression coefficients can be accomplished by multiplying all the nonzero values for the indicator codes by -1; in other words, turn the 1s to -1s, as in Table 3 which has the no protest group as the reference. This can be done by programming the matrix, as in the prior example.

Table 3.

	X1	X2
No protest (`protest` = 0)	0	0
Individual protest (`protest` = 1)	-1	0
Collective protest (`protest` = 2)	0	-1

The PROCESS command that accomplishes this is (using the `protest` variable in the data file, which has the groups coded no protest = 0, individual protest = 1, collective protest = 2).

```
process y=liking/x=protest/model=0/mcx=5/xcatcode=0,0,-1,0,0,-1.
```

```
%process (data=protest,y=liking,x=protest,mcx=5,xcatcode=0 0 -1 0 0 -1)
```

```
process (data=protest,y="liking",x="protest",model=0,mcx=5,xcatcode=c(0,0,-1,0,0,-1))
```

The output below results

```
Model  : 0
    Y  : liking
    X  : protest

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest      X1      X2
    .000    .000    .000
   1.000  -1.000    .000
   2.000    .000  -1.000

**************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.3102      .1614    32.9083      .0000     4.9909     5.6296
X1           -.5158      .2255    -2.2870      .0239     -.9621     -.0695
X2           -.4431      .2231    -1.9863      .0492     -.8845     -.0016
```

As you can see, this output is largely identical to the output when using "1" rather than "-1" for the indicator codes from earlier in this document. However, the signs of the regression coefficients (and the endpoints of the confidence interval) have reversed. The regression coefficient for X1 is now the no protest mean minus the individual protest mean (-0.516 = 5.310 – 5.826) and the regression coefficient for X2 is now the no protest mean minus the individual protest mean (-0.443 = 5.310 – 5.753). The test of equality of the three means is unaffected, $F(2,126) = 3.055$, $p = .051$, and the model still reproduces the group means:

No protest group: $\hat{Y} = 5.310 - 0.516(0) - 0.443(0) = 5.310$

Individual protest group: $\hat{Y} = 5.310 - 0.516(-1) - 0.443(0) = 5.826$

Collective protest group: $\hat{Y} = 5.310 - 0.516(0) - 0.443(-1) = 5.753$

For more information on the coding systems for multicategorical variables available in PROCESS and programming your own system for representing groups, see the PROCESS documentation in *Introduction to Mediation, Moderation, and Conditional Process Analysis*.