Teaching How to Plan for Creating the Variables You Need from the Variables You Have

Jane E. Miller, Ph.D.

INTRODUCTION

For many statistical analyses, variables available in an existing dataset are not yet in the form needed to address the research question of interest. Examples of situations in which new variables need to be created include:

- A student wants to study total family income, but the dataset has separate variables measuring income components such as earned income, government benefits, and alimony.
- Someone wants to study respondent's eligibility for social programs such as Medicaid that are based on multiples of the U.S. Federal Poverty Level (Centers for Medicare and Medicaid Services 2013), but the dataset reports income in dollars and has separate measures of number of children and adults needed to identify the threshold for the size and age composition of the household.
- A researcher is interested in comparing outcomes for children, working age adults, and the elderly, but the dataset reports continuous age in single year units.

Sometimes it will not be possible to create the desired form of a variable from the variables in a dataset, such as:

- Calculating Federal Poverty Level from family income, if no measure of household size or age composition is available.
- Calculating age in single years from age group of respondent, because it is impossible to retrieve the more detailed age from a less detailed (categorical) measure.

Some statistics classes teach how to use statistical software menus or syntax to create new variables, usually working from a set of instructions such as the cut-offs needed to create classifications, or formulas needed to compute a new variable. In this paper, I demonstrate how to back the process up a couple of steps so that students learn (1) *why* planning steps are necessary to get from the variables they have to the variables they need for their analysis, (2) *what information* they need to identify the new variables they need for their topic; and (3) *how* to write clear instructions on how to get from the variables available in their dataset to the variables they need for their analysis. These planning skills are critical for students as they learn to conceptualize and measure variables to match their own research questions, and will help them anticipate how to write about variable transformation in their research papers.

In the research methods and statistics courses that I teach, I require students to conduct simple statistical analyses of data from publicly available datasets. Having taught those courses for many years, I have learned the importance of separating the planning of how to create new variables from using statistical software to create those variables in a database. In the classroom, we learn the conceptual basis of the tasks and conduct the planning away from the temptation to start clicking away in the statistical software without forethought about what to do and why. In the subsequent computer lab session, students use their planning notes to guide them through the point-and-click steps of creating new variables in their dataset. The idea is to create a "recipe" of typed notes that someone who is unfamiliar with the concepts or data could follow if they know how to use the software and have access to the dataset. The original variables are the ingredients, the steps are the cooking techniques, and the new variables are the finished dish.

I teach the skills in that sequence because planning and implementing creation of new variables involves distinct steps and skill sets, each of which is new and challenging to many students. Also, whereas point-and-click instructions or software lessons are limited to a particular topic and data, learning these planning steps teaches students how to approach the process for other research questions they will study later in their career. The aim is to help students master the abstract concepts behind the planning

steps by applying them to several concrete examples so they will be able to transfer the skills to other related tasks (Willingham 2009).

In the next three sections, I describe the steps involved in planning for data preparation: (1) Identifying the source variables, (2) becoming familiar with how variables on their topics are treated in the literature, and (3) writing the logic or math needed to correctly create the new variables. I then illustrate the planning process, tracing a specific topic example through each of the steps, and providing exercises on other topics to reinforce the concepts. These planning steps should be undertaken with pencil and paper or in a word processor document before students attempt to create the new variables in statistical software. In other words, teach the students to write the recipe before they attempt to cook an unfamiliar dish!

IDENTIFYING THE PERTINENT SOURCE VARIABLES IN THE DATASET

The first step in the planning process is for students to identify the original or source variables on their topic that are *available in the dataset they will use*, because the content, units, and coding of those variables will determine what new variables can be created, as explained above. That information can be gleaned from a codebook for that dataset and should be confirmed by examining the variables in the electronic database itself (Miller 2013).

FINDING REFERENCES ON STANDARD WAYS OF HANDLING CONCEPTS

The next key background step is anticipating what the new variables should look like and why. To do so, students should learn to find authoritative references on how the concepts they are studying are usually handled in the literature, including criteria for identifying credible reference sources. A university library web site or other authoritative source can provide pertinent guidelines (Yale College Writing Center 2013). In my experience, if students lack substantive context for the concepts they are studying and an explanation of why such background is important, they tend to make up criteria for a new variable off the top of their head, at best reinventing the wheel, and at worst devising some arbitrary, facile approach (such as dividing by 10) just to complete the assignment.

For some research questions, students need to learn the pertinent program criteria or clinical thresholds such as the eligibility threshold for a social program, or ranges of blood pressure used to identify hypertension. For other research questions, they should become acquainted with empirical approaches such as classifying distributions into percentiles (widely used for topics like standardized test scores) or taking logs of variables such as income, for example. Sometimes, new variables must be created due to something about the student's dataset, such as when some subgroups are too small to be analyzed separately – a problem that can be overcome by creating a new version of the variable that combines several small categories, e.g., a three-category race variable instead of a five-category one.

WRITING THE INSTRUCTIONS FOR CREATING NEW VARIABLES

Planning how to create a new variable from existing ones uses a combination of logic and mathematics, depending on the levels of measurement of the original and new variables. See Chambliss and Schutt (2012) for more on levels of measurement, also known as “types of variables." The logic and mathematical steps can be thought of as word problems, converting the information learned in the previous step about how the concepts for the original and new variables relate to one another into formulas or instructions on how to classify values. Readers who are familiar with the state of quantitative (il)literacy among many students today (Paulos 1988; Steen 2001) and many students' aversion to word problems will instantly anticipate the challenge of this step and therefore the necessity to devote adequate time and practice in class and homework.

The nature of the "word problem" differs depending on the type(s) of variables. Planning how to create a new continuous variable involves writing a formula, whereas planning a new categorical variable involves classification instructions such as those conveyed using a grid like that shown below.

New continuous variables

To create a new continuous variable from existing continuous variables involves arithmetic or statistical computations that can be written using a *formula*. Continuous variables can only be created from other continuous variables or from a combination of continuous and categorical variables.

- Some of these computations will be quite simple, involving only one variable and basic operations such as dividing a variable by a constant or taking logarithms.
- Some computations are slightly more complex, involving two or more variables or more complicated arithmetic such as calculating miles per gallon from mileage and money spent on gasoline, or body mass index from height and weight.
- The most complicated computations involve multiple standards and/or arrays, such as calculating income as a multiple of the Federal Poverty Level, which must take into account family size and age composition (U.S. Census Bureau 2013).

For continuous variables, units must be specified for both the original and new variables.

In addition, stress the importance of paying attention to the units of measurement in the original variables, for thresholds, and in formulas to ensure that the calculations and comparisons are done correctly. For example, if height and weight were reported in British units (feet and inches; pounds) in the original data, additional steps would be needed to convert them into meters and kilograms, respectively, before they could be used in the standard formula to compute body mass index and compared against the WHO thresholds for weight status.

New categorical variables

To create a new categorical variable involves showing how each value of the existing variable is *classified* into values of the new variable. The original variables can be either continuous or categorical.

- Continuous age can be classified into age groups.
- A detailed categorical variable can be simplified into a less detailed one, such as collapsing 5-year age groups into 10-year age groups.
- Two variables such as race and ethnicity can be cross-tabulated to create a single variable, such as non-Hispanic white, non-Hispanic black, Hispanic, and other.

For categorical variables, values (also known as "codes") and value labels will also need to be devised and recorded.

For either level of measurement, the planning instructions should show how every possible value of the original variable or variables maps into a value of the new variable. For example, if age in years is known for every respondent in the sample, the new age group variable should also have a value for every respondent.[1] If a respondent was missing information on the original variable, that same respondent should also have a missing value on the new variable.

In SPSS software, creating a new continuous variable from one or more existing continuous variables is done with the "compute" procedure, whereas creating a new categorical variable is done with the "recode into new variables" procedure — a distinction that can help students anticipate whether they need to write a formula or classification instructions.

EXAMPLE PLANNING EXERCISE

To illustrate the exercise, I will trace the example of the variables needed for an analysis of obesity, working from separate measures of height and weight. Respondents' body weight alone cannot be

[1] Care must be taken to create mutually exclusive and exhaustive values of the new variable, defining one and only one value of the new variable for each possible value of the original single-response variable or combination of values of two cross-tabulated variables. See Chambliss and Schutt (2012) and Miller (2015) for more on well-defined categorical variables and single- and multiple-response items.

used to assess whether each respondent is obese because whether a certain weight implies obesity also depends on the person's height: A 140 lb. adult who is 5'10" tall would be severely underweight, whereas someone at the same weight who is 5'0" tall would be obese. To take into account both height and weight, a measure known as body mass index (BMI) is used to assess obesity, calculated from continuous measures of weight and height.

The planning process for this particular example involves several steps, each of which should be explained and illustrated during the in-class demonstration.

1. Naming the original (source) variables to be used in the creation of the new variables, in this example HGHTMTR and WGHTKG. For each variable, specify the variable name (acronym in the dataset), variable label, and missing values, if any.
    a. For source variables that are continuous, the label should specify units of measurement.
    b. For source variables that are categorical, numeric codes and labels should also be specified for each category.
2. Computing body mass index (BMI), which is a simple index of weight-for-height that is commonly used to classify underweight, overweight and obesity in adults. It is defined as the weight in kilograms divided by the square of the height in meters (kg/m^2) (WHO 2013). The new variable will need a name, label, units, and notes about missing values.
3. Creating a multi-category indicator of weight status category from BMI using standard ranges provided by the World Health Organization (2013). Underweight is defined as BMI<18.50, normal weight: BMI 18.50-24.99, overweight BMI 25.00-29.99, and obese as BMI 30 or higher. The new categorical variable will need a name, label, values (codes for each value), category labels (labels for each value), and notes about missing values.
4. Creating a binary indicator (obese or not) from the categorical weight status indicator. The new categorical variable will need a name, label, values (codes for each category), category labels (labels for each value), and notes about missing values.

I intentionally chose a complex, multi-step process for the in-class demonstration because it provides several opportunities for students to observe new variables being created from existing ones, and to show creation of new continuous, detailed categorical, and new simplified categorical variables.

Planning notes for step 1:

List the information about the original variables:

Variable name	Variable label	Missing values	Level of measurement
HGHTMTR	height in meters	999	Continuous variable
WGHTKG	weight in kilograms	999	Continuous variable

Planning notes for step 2:

Write the name, label, units, and missing values for the new continuous variable to be created:

BMI = "body mass index, (kg/m^2)" Missing values = 999.

Write the formula for computation of new continuous variable from two continuous source variables:

BMI = WGHTKG / (HGHTMTR^2), where "^" indicates exponentiation

Write notes about the situations in which the new variable will have missing values:

IF WGHTKG = 999 or HGHTMTR = 999, THEN BMI = 999

Planning notes for step 3:
Create a grid to show how each value of the continuous BMI variable maps into values of the new categorical variable BMICAT. The value ranges of the original variable and their classifications for the new variable are derived from the WHO cutoffs mentioned above.

NAME of original variable _BMI_ **LABEL for original variable** __Body mass index, (kg/m^2)__	**NAME of new variable** __BMICAT__ **LABEL for new variable** __Weight status category____	
Values of original variable	**Values (codes) for new variable**	**Value labels of new variable**
<18.50	1	Underweight
18.50 - 24.99	2	Normal range
25.00 - 29.99	3	Overweight
≥30.00	4	Obese
999	9	Missing

Planning notes for step 4:
Create a grid to show how each value of the categorical variable BMICAT maps into values of the new categorical variable OBESE.

NAME of original variable _BMICAT_ **LABEL for original variable** __ Weight status category__	**NAME of new variable** _OBESE__ **LABEL for new variable** __ Obesity indicator____	
Values of original variable	**Values (codes) for new variable**	**Value labels of new variable**
1, 2, and 3	0	Non-obese
4	1	Obese
999	9	Missing

PEDAGOGICAL APPROACH

I recommend that this material be taught using the following sequence of activities: (1) demonstrate the BMI example as part of lecture; (2) walk the students through another topic soliciting input from them at each step; (3) break the class into groups to practice the steps on an assigned topic; and (4) in class, explain the instructions for a related homework assignment (see instructions below) so students can ask questions before heading out to tackle it on their own. After steps 1 through 3, ask the class, "are you confident that you could go home and complete this series of steps for your own research question and data? If not, which steps would you like me to explain again? Which ones would you like to practice again?" If they would like additional in-class practice, solicit new topics or questions from the students.

SUMMARY

This exercise prepares students to understand why it is often necessary to create new variables before they can conduct statistical analyses to address their research question using existing data. It also provides a systematic approach for planning creation of new variables, including identifying source variables, finding relevant references about how such variables are conventionally analyzed, becoming familiar with units or categories of the original and new variables, and writing formulas or classification instructions to create the new variables from the original variables.

REFERENCES

Centers for Medicare and Medicaid Services. 2013. "Medicaid Eligibility." Available online at http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Eligibility/Eligibility.html. Accessed January 2014.

Chambliss, D. F., and Schutt, R. K. 2012. M*aking Sense of the Social World: Methods of Investigation, 4th Edition*. Thousand Oaks, CA: Sage Publications.

Miller, J.E. 2015. *The Chicago Guide to Writing about Numbers, 2nd Edition*. Chicago: University of Chicago Press.

-----. 2013. "Getting to know your variables: The foundation for a good working relationship with your data." In *2013 Proceedings of the American Statistical Association, Statistical Computing Section*. Alexandria, VA: American Statistical Association, pp. 2016-27.

Paulos, J.A. 1989. *Innumeracy: Mathematical Illiteracy and Its Consequences.* New York: Farrar, Straus, and Giroux.

Steen, L.A. 2001. *Mathematics and Democracy: The Case for Quantitative Literacy.* Princeton NJ: Woodrow Wilson National Foundation.

U.S. Census Bureau, Social, Economic, and Housing Statistics Division. 2013. Poverty Thresholds. Available online at http://www.census.gov/hhes/www/poverty/data/threshld/index.html. Accessed January 2014.

Willingham, D.T. 2009. "Why is it So Hard for Students to Understand Abstract Ideas?" In *Why Don't Students Like School? A Cognitive Scientist Answers Questions about How the Mind Works and What it Means for the Classroom.* San Francisco: Jossey-Bass.

World Health Organization. 2013. "BMI Classification." Available online at http://apps.who.int/bmi/index.jsp?introPage=intro_3.html. Accessed January 2014.

Yale College Writing Center. 2013. "Scholarly vs. Popular Sources." Available online at http://writing.yalecollege.yale.edu/scholarly-vs-popular-sources. Accessed January 2014.

IN-CLASS EXERCISES:

Divide the students into small groups. Assign each group one of the following topics and ask them to outline the steps needed to get from the original variables available in the hypothetical dataset to variables needed to test the hypothesis or answer the word problem. It may help to provide numeric values of the original variables for a few cases for students to test out their formulas and classification instructions, but their planning instructions should be written so as to be applicable for *any* plausible hypothetical input values of those variables.

1. Hypothesis: "Households with more than six members are more likely than smaller families to earn incomes less than $20,000 per year."
    a. Original variables
        i. Number of adults in the household (continuous)
        ii. Number of children in the household (continuous)
        iii. Annual family income (continuous)

    b. New variables
        i. Total household size, calculated from # adults and # children (continuous)
        ii. Total family size, classified into <=6 versus 6+ (categorical)
        iii. Total family income, classified into <$20K versus $20K+ (categorical)

2. Calculate gas mileage, cost per mile driven, and total trip cost for a multi-stop trip.
    a. Original variables
        i. Amount of gas purchased at each stop (continuous)
        ii. Cost of gas purchased at each stop (continuous)
        iii. Number of miles driven between stops (continuous)

    b. New variables
        i. Gas mileage, calculated from gallons of gas purchased and # miles driven since the previous stop (continuous)
        ii. Cost per mile, calculated from cost of gas and miles driven since the previous stop (continuous)
        iii. Total trip cost, calculated from cost of gas purchased at all stops (continuous)

Have groups exchange planning notes with one another, to assess whether those notes provide adequate guidance on the steps needed to create the new variables based solely on the information provided in those notes.

NOTE: IF STUDENTS DO NOT HAVE THEIR OWN RESEARCH QUESTIONS, THE ABOVE TOPICS CAN BE USED AS A HOMEWORK ASSIGNMENT

HOMEWORK ASSIGNMENT - IF STUDENTS HAVE THEIR OWN RESEARCH QUESTIONS AND DATA

Before you analyze data for your research question, think about (and discuss with your professor or research mentor) whether you need to create any new variables in order to address that research question using an existing dataset available to you in electronic form. Examples include:

- categorical versions of continuous variables, such as age group from age in years
- simplified (collapsed) categorical variables, such as a 3 racial/ethnic groups from a detailed list of many racial/ethnic groups;
- a binary indicator such as low birth weight ("yes" versus "no") from a continuous measure of birth weight;
- a new continuous variable such as Body Mass Index from height and weight

This planning step should be done *before* you point and click (or write syntax) to create one or more new variables in your electronic database.

Create a Word document that includes the following information:

1) The hypothesis for your research project in which you name the independent and dependent variables needed to answer that question (in the form you wish to analyze them).

2) List the variables in the original data set that measure the concepts you plan to study, along with the following information about each of these variables, which should be available in the documentation for the data set:
    a) Variable name (acronym);
    b) Variable label – a short phrase describing what the variable measures;
    c) Units or categories.
    d) Missing value codes, if any.

    NOTE: The names, labels, units, coding, and missing values of the original variables must match those in the actual dataset that you will use for your statistical analysis. Do not make up hypothetical versions!

For each new variable you will create, specify:

3) A variable name (acronym). It should convey the content (meaning) of the new variable, and if dates or survey rounds pertain, an abbreviation for that information.

4) A label – a short phrase describing what that variable measures, including units where pertinent.

5) Missing value codes. If there were missing values on the original variable(s), those should be mapped into missing values on the new variable.

6) Explain the logic and math that will be used to create the new variables. Do not write the syntax or describe the point-and-click steps you will use in your software program. To learn what formulas, cutoffs or standards pertain to your topic, review the literature in your field.

7) For new *continuous* variables,
    a) Write the mathematical formula by which that new variable is created from one or more variables available in the original dataset.
    b) Specify the units of the original variable(s) and the new variable.

(Continued on next page)

8) For new *categorical* variables,
   a) Fill out the attached planning grid, making sure to show how every possible value of the original variable maps into a value of the new variable.

   b) List the:
      i. Code (numeric value) that the new variable will take on for each value or set of values of the original variable;

      ii. Value label (descriptive phrase) for each value (category) of the new variable.

   HINT: It often helps to write the value labels for categories of the new variables first, then give each category a numeric code, and finally list the values of the original variable that map into each category.

9) Provide a bibliographical citation for each reference source you used to identify formulas or classification criteria, and cite that reference next to the pertinent step in the planning process.

PLANNING TEMPLATE FOR NEW CATEGORICAL VARIABLES

NAME of original variable _____ **LABEL for original variable _____**	**NAME of new variable _____** **LABEL for new variable _____**	
Values of original variable	**Values (codes) of new variable**	**Value labels**

Citation(s) for source(s) of information about how to create the new variable, e.g., formula for calculation of a continuous variable, or cutoffs for classification of new categorical variable.